UNDERSTANDING THE MOTION OF A HUMAN STATE IN VIDEO CLASSIFICATION

A Defense
Presented to
The Academic Faculty

By

Daniel Alejandro Castro Chin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology

May 2019

UNDERSTANDING THE MOTION OF A HUMAN STATE IN VIDEO CLASSIFICATION

Approved by:

Dr. Irfan Essa, Advisor
School of Interactive Computing, College of Computing
Georgia Institute of Technology

Dr. Dhruv Batra
School of Interactive Computing, College of Computing
Georgia Institute of Technology

Dr. James Hays
School of Interactive Computing, College of Computing
Georgia Institute of Technology

Dr. Devi Parikh
School of Interactive Computing, College of Computing
Georgia Institute of Technology

Dr. Rahul Sukthankar
Google Research
Alphabet

Date Approved: March 25, 2019

No te dejes confundir. Busca el fondo y su razón. Recuerda, se ven las caras... pero nunca el corazón.

Rubén Blades

ACKNOWLEDGEMENTS

This work would have not been possible without the drive and work ethic I was taught by my family (Mom, Dad, Bea, Pipo, gracias) and my formal education: the International School of Panamá and the Georgia Institute of Technology. I am incredibly thankful for all of my friends and colleagues whom I have had the pleasure of working with. Most importantly, I am truly grateful for the guidance year after year from my advisor, Dr. Irfan Essa and the guidance of my committee members, Dr. Dhruv Batra, Dr. James Hays, Dr. Devi Parikh and Dr. Rahul Sukthankar, thank you so much for your precious time. I am especially thankful for the trust Dr. Essa has had in me and the work that he put in over the years to shape my academic career – from all the way back to the first time we met when I took his undergraduate class in Computational Photography in Barcelona, Spain. Special thanks to my love, Carley-Beth, for all of your support to push me to this sometimes mythical finish line. To each and every one of my collaborators, Dr. Matthias Grundmann, Dr. Vivek Kwatra, Dr. Vinay Bettadapura, Dr. Edison Thomaz, Dr. Gregory Abowd, Dr. Henrik Christensen, Dr. Amy Bruckman, Dr. Eric Gilbert, Dr. Rosa Arriaga, Dr. James Hays, Steven Hickson, Aneeq Zia, Patsorn Sangkloy, Hayley Evans, Marisol Wong-Villacres, Michaelanne Dye, and all of my collaborating undergraduate and Masters students, Dhruv Mehra, Sean Dai, Wataru Ueno, Shaohui Xu, Alexander Neal, Patrick Violette, Sam Skinner, Yushi Sun, Luis Perez, Michael Maurer, Vikram Jain and Bhavishya Mittal, thank you. And of course, to my entire research lab, thank you all so much. Lastly, thanks to the country that raised me, Panamá, for teaching me to always approach life with a smile, to be grateful and to never give up, for we are capable of everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

For the last 50 years we have studied the correspondence between human motion and the action or goal they are attempting to accomplish. Humans themselves subconsciously learn subtle cues about other individuals that gives them insight into their motivation and overall sincerity. In contrast, computers require significant guidance in order to correctly determine deceivingly basic activities. Due to the recent advent of deep learning, many algorithms do not make explicit use of motion parameters to categorize these activities. With the recent advent of widespread video recording and the sheer amount of video data being stored, the ability to study human motion has never been more essential. *In this thesis, we propose that our understanding of human motion representations and its context can be leveraged for more effective action classification.*

We explore two distinct approaches for understanding human motion in video. Our first approach for classifying human activities is within an egocentric context. In this approach frames are captured every minute at a low frame rate video that represents a summary of a persons' day. The challenge in this context is that you do not have an explicitly visual representation of a human. In order to tackle this problem we therefore leverage contextual information alongside the image data to improve the understanding of our daily activities. In this approach, motion is implicitly represented in the image data given that we do not have a visual representation of a human pose. We combine existing neural network models with contextual information using a process we label a late-fusion ensemble. We rely on the convolutional network to encode high-level motion parameters which we later demonstrate performs comparably to explicitly encoding motion representations such as optical flow. We also demonstrate that our model extends to other participants with only two days of additional training data. This work enabled us to understand the importance of leveraging context through parameterization for learning human activities.

In our second approach, we improve this encoding by learning from three represen-

tations that attempt to integrate motion parameters into video categorization: (1): regular video frames (2): optical flow and (3): human pose representation. Regular video frames are most commonly used in video analysis on a per-frame basis due to the nature of most video categories. We introduce a technique which enables us to combine contextual features with a traditional neural network to improve the classification of human actions in egocentric video. Then, we introduce a dataset focused on humans performing various dances, an activity which inherently requires its motion to be identified. We discuss the value and relevance of this dataset along the most commonly used video datasets and among a handful of recently released datasets which are relevant to human motion. Next, we analyze the performance of existing algorithms with each of the motion parameterizations mentioned above. This assists us in understanding the intrinsic value of each representation and a better understanding of each algorithm. Following this, we introduce an approach that utilizes each of the motion parameterizations concurrently, in order to have a better understanding of the video. From here, we introduce a method to represent a human pose over time to improve human video categorization. Specifically, we look at specific joint distances over time to generate features that represents the distribution of specific human poses over time. Performance of each individual metric will be computed and analyzed in order to assess their intrinsic value. The main objective and contribution of our work is to introduce a parameterization of human poses which improve action recognition in video.

CHAPTER 1
INTRODUCTION

The field of computer vision has spent its entire existence attempting to understand, replicate or improve the abilities of the human visual system. Humans have the remarkable ability to understand the dynamic world around them without being constantly overwhelmed. We have an incredible sense of motion. We can estimate relative speeds while in a moving car, we have the ability to predict the location of a tennis ball moving upwards of 100 miles per hour and react in real-time and we can even draw correlations between how a person walks and the progression of certain cognitive disorders [1] [2]. All of these areas are existing problems in the field of computer vision which we strive to solve. Even without human subjects, motion cues in video can be used to predict the camera's path [3], and has been successfully leveraged in the stabilization of videos [4] [5] [6]. The problems in the field are vast, and their attempted solutions admirably detailed and extensive. We have discovered these cues because of our inherent desire to understand what is it exactly that goes on in our minds, and more specifically, in our sight.

1.1 Importance of Motion & Movements in Video

Motion plays a vital role in the understanding of a scene. We define motion in this work as any change which occurs in a natural scene. A scene in this context is defined as a view of a real-world environment. We refer to movements as any changes in the state of the scene. A state is a specific moment in a scene. Lastly, an activity takes place in a particular state. Therefore, movements tend to produce motion which are often captured in an image and therefore encode part of the state of the scene. In this work we will highlight how convolutional networks are able to encode motion in cases where it is embedded in the image frame and how explicitly incorporating the parameterization of motion can show

improvements in video classification.

1.1.1 Egocentric Motion

In the egocentric image space, images are captured passively by a device that is generally attached to the subject. The movement of the camera is therefore often captured by the camera depending on what activity they are conducting. Egocentric motion has been used previously in detecting actions [7] [8], predicting gaze and attention [9] [10] and improving object recognition [11]. If the subject is walking or cycling, the image frames are likely to encode a blurrier shot because of the subjects' movement. In contrast, sitting down whilst working is likely to generate a fairly clear image, encoding zero motion. This encoded motion is vital in distinguishing activities that have a lot of movement from those that don't in our work in recognizing daily activities.

1.1.2 Motion of a Human Pose

The problem of inferring and understanding motion in a human pose can most famously be tracked back to the experiments conducted by psychologist Gunnar Johansson. [12]. In a follow-up to his seminal work, written for *Scientific American*, he presents a figure of two individuals dancing in the dark [13]. These images were put together by having participants wear highly reflective tape in the darkness and shining a light to simulate what is trivially generated today by pose detection in computer vision. Johansson sought to understand the minimal amount of information required for our senses to recognize the underlying subject and more specifically, their biological motion. With a mere 10 points in a black image, he was able to represent a human in motion. We present a snippet of this original work in Figure 1.1, where we can appreciate the work done to extract this information. Computationally, this is the equivalent of compressing an entire image down to ten pixels and still retain the underlying high-level action in the scene. What we don't understand however is how computers perform in this particular experiment and whether

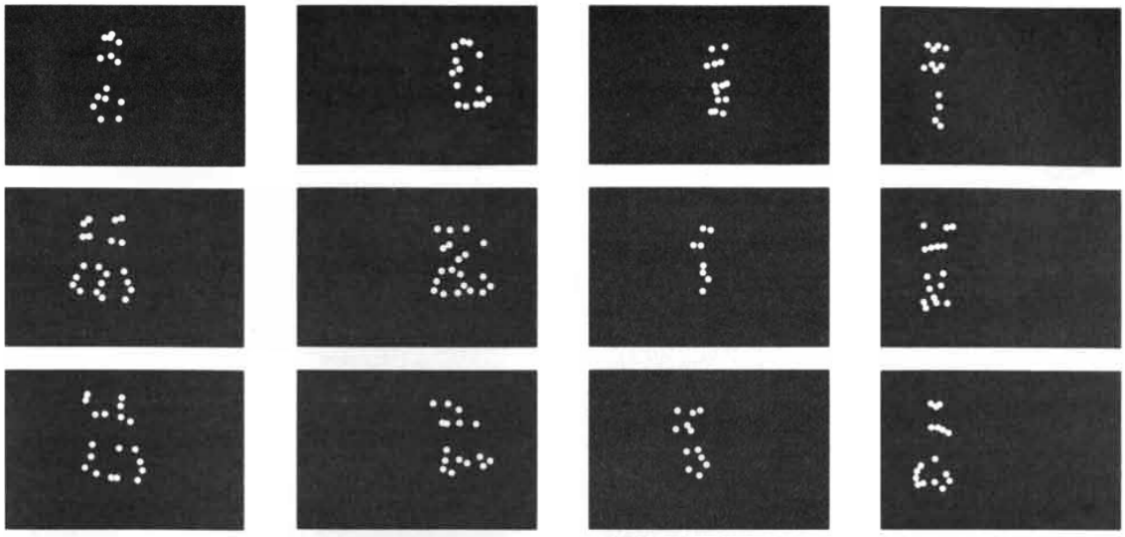

Figure 1.1: Here we can appreciate a subset of the frames published by Johansson [13] as he explored the intricacies of human motion in dance.

they can perform equally well with such constrained data. A bulk of our work will focus on understanding these constraints, testing our learning within them and understanding novel ways of learning given those parameters.

In this work, we seek to understand the level of detail that computers need to classify a particular action. In the last 10 years, there has been a growing tendency of simply passing in all of the information we have available into an algorithm (most recently a neural network) with the intuition that it would filter what is relevant and discard the rest. This tendency was most famously introduced by Khrizhevsky et. al [14] in their seminal ImageNet work. Although these techniques perform surprisingly well in image classification, it remains unclear whether it is reasonable to extend these techniques to video and even less so whether they extend to action recognition in a sensible manner. This neural approach is merely common because of the intuition that a video is inherently a collection of images. This generalization falls apart when taking into account the fact that the change between image frames intrinsically contains motion.

Reasonably, this difference between two frames has been looked at extensively in the field of computer vision. Grundmann et. al. leveraged the motion of specific features in a

video to infer the camera's path [6]. Sheikh et. al. used this difference to compute complex background subtraction in freely moving cameras [15]. The entire field of computing optical flow is based on the changes between frames that can help encode motion into a representation that computers can better understand. Therefore, the main goal is to approach the problem with insights from how humans understand a pose image (as shown in Figure 1.1).

1.2 Motion in Humans

In this work we explore the problem of recognizing the actions of humans in images and video. We wish to extend the definition of an action presented by Poppe [16] in his survey of action recognition. This definition identifies an action as a set of action primitives that encompass a whole-body movement. They define an action primitive as an "atomic movement that can be described at the limb level". We further this by defining a dynamic action as a set of action primitives that require movement to be identified (i.e. such as a specific dance), meaning that they are inherently difficult or impossible to identify using a single snapshot of their pose.

Although it is fairly obvious that motion plays a key role in humans understanding the actions of others it is not entirely clear what feature representations we use to actually understand those actions. In order to further our understanding, we focused a significant portion of our analysis on humans dancing [17]. In this work we explored 1000 videos of humans dancing to 10 different types of dances. The goal of this work was to emphasize the importance of motion cues and deemphasize the relevance of visual ones. One of the key problems we had found with previous work in the area was the overwhelming amount of datasets that could be solved using a single carefully selected frame and visual cues. In order to mediate this issue, we presented a dataset comprised solely of dynamic actions and tested common algorithms on the visual data and pose representations which were stripped of traditional visual cues. Further, we sought to understand which existing

learning algorithms would be best suited for understanding and learning motion cues.

Given these contexts, the importance of motion is the focus of our work. We therefore formulate the thesis statement as follows: **Our understanding of human motion representations and its context can be leveraged for more effective action classification.** In the following chapters we will cover prior work that has established the foundation for this thesis statement, and detail our proposed work for the completion of the thesis.

Chapter 2 will cover work we conducted in recognizing daily activities from egocentric images [18]. This work focused on analyzing images taken from a wearable camera processed with relevant contextual information in order to improve the standard convolutional neural network classification. Through the use of an ensemble method we successfully improve standard approaches by fusing intuitive features with convolutional features in order to improve the overall performance of classifying the activity that a person is conducting.

In Chapter 3 we will overview the motivation behind the creation of the "Let's Dance Dataset", and the details of this work. Additionally, we will discuss the different machine learning approaches we analyzed in three different data representations (RGB, Optical Flow, and a Pose Visualization) in order to assess the current performance of state-of-the-art methods. We will also discuss new datasets that were introduced alongside our work that further motivate the relevance of dynamic actions in the field of action recognition. We will highlight the importance of data representations when dealing with human actions and how to better understand the motion of a human pose.

In Chapter 4 we will present our work on the parameterization of human poses for video classification. After exploring motion from two camera perspectives (egocentric, standard), we propose taking a deeper look at how human beings process and understand the motion of a pose representation. We will use the work of Johansson [12] as a foundation for how much a human can understand from different pose representations and what components are particularly necessary for detecting an action. We will conduct high-level user studies to understand if a human can identify basic actions from a short clip of the visualized

joints. We will then use these insights to understand what is the best way to represent our data and which machine learning methods are best suited for classifying these tasks. We will apply these insights to both the egocentric and human context to demonstrate the benefit of our proposed parameterization. We will demonstrate that our approach to parameterizing a human pose can significantly improve the performance of standard neural network by combining them using late-fusion approaches that we established in Chapter 2. Specifically, we will develop a hashing algorithm for a human pose that encodes the distance between specific limbs in order to characterize the current pose. A histogram of these poses will then be used to characterize each specific dance. Using late-fusion we will be able to combine human poses with standard images for significant classification improvements.

CHAPTER 2

CLASSIFYING HUMAN ACTIVITIES FROM EGOCENTRIC IMAGES

We present an approach for classifying human activities taken from a passive egocentric camera. We leverage both image and contextual information in our approach to improve our overall classification by integrating relevant features into a convolutional learning method [18]. We leverage existing deep convolutional networks to detect intrinsic motion features in an image and combine their predictions with the parameterization of contextual features (time and color histograms) to attain the best classification. Additionally, we demonstrate promising results with fine-tuning the initial model to two unseen participants with just a single day of training data.

This work was presented at the 2015 ACM International Symposium on Wearable Computers [18].

2.1 Introduction

The ability to automatically monitor and infer human behaviors in naturalistic environments is essential for a wide range of applications in areas such as context-aware personal assistance, healthcare, and energy management. Recently, wearable egocentric cameras such as the GoPro [1] and Narrative [2] have become ubiquitous, enabling a new form of capturing human experience. Egocentric photos taken by these cameras can provide rich and objective evidence of a person's everyday activities. As a result, this data collection approach has been extensively used in a variety of research domains, particularly in healthcare. Health-related applications that have leveraged first-person photos include, but are not limited to, retrospective memory support [19], dietary assessment [20, 21], autism sup-

[1] http://www.gopro.com
[2] http://www.getnarrative.com

port [22], travel and sedentary behavior assessment [23, 24], and recognition of physical activities [25].

Figure 2.1: Example images from the Daily Activities Dataset of 40,000 egocentric images with their respective labels. The classes are representative of the number of images per class for the dataset. Note: We handpicked family images for this figure so they did not contain family subjects (for privacy and anonymity concerns).

Outside of privacy, first-person photo capture with wearable cameras has one additional serious challenge. Once photographs have been taken it is necessary to review them to identify moments and activities of interest, and possibly to remove privacy-sensitive images. This is particularly challenging when wearable cameras are programmed to take snapshots periodically, for example: every 30 or 60 seconds. At this rate, thousands of images are captured every week, making it imperative to automate and personalize the process of image analysis and categorization. However, standard motion-driven approaches to classifying video often fall apart with approaches that have very low frame rate. Therefore, the frequency at which we capture snapshots has to carefully balance capturing just the right amount of information to be able to learn from.

We describe a computational method leveraging state-of-the-art methodologies in machine learning to automatically learn a person's behavioral routines and predict daily activities from first-person photos and contextual metadata such as the day of the week and

the time of day. Example of our daily activities include cooking, eating, watching TV, working, spending time with family, and driving (see Table 2.1 for a full list). The ability to objectively track such daily activities and lifestyle behaviors is extremely valuable since behavioral choices have strong links to chronic diseases [26].

To test and evaluate our method, we compiled a dataset of 40,103 images representing everyday human activities. The dataset has 19 categories of activities and were collected by one individual over a period of six months "in the wild". Given the egocentric image and the contextual date-time information, our method achieves an overall accuracy of 83.07% at determining which one of these 19 activities the user is performing at any moment.

Our classification method uses a combination of a Convolutional Neural Network (CNN) and a Random Decision Forest (RDF), using what we refer to as a CNN late-fusion ensemble. It is designed to work on single images captured over a regular interval as opposed to video clips. Capturing hours of egocentric video footage would require tethered power and large storage bandwidth, which still remains impractical. An example of our input egocentric image and the output class prediction probabilities is shown in Figure 2.1. In this section of our work we accomplished:

- A robust framework for the collection and annotation of egocentric images of daily activities from a wearable camera.
- A CNN+RDF late-fusion ensemble that reduces overfitting and allows for the inclusion of local image features, global image features, and contextual metadata such as day of the week and time.
- A promising approach to generalize and fine-tune the trained model to other users with a minimal amount of data and annotation by the user. We also get insights into the amount of data the first user needs to collect to train a classifier and how much data subsequent users need to collect to fine-tune the classifier to their lifestyle.
- A unique dataset of annotated egocentric images spanning a 6 month period and a

CNN+RDF late-fusion ensemble model fit to that data.

2.2 Related Work

Activity Analysis: Discovering daily routines in human behavior from sensor data has been an active area of research. With a dataset of 46 days of GPS sensor data collected from 30 volunteer subjects, Biagioni and Krumm demonstrated an algorithm that uses location traces to assess the similarity of a person's days [27]. Blanke and Schiele explored the recognition of daily routines through low-level activity spotting, with precision and recall results in the range of 80% to 90% [28]. Other proposed techniques for human activity discovery have included non-parametric approaches [29], and topic modeling [30].

One of the most comprehensive computer-mediated analysis of human behaviors in naturalistic settings was done by Eagle and Pentland [31]. By collecting data from 100 mobile phones over a 9-month period, they were able to recognize social patterns in daily user activity, infer relationships, identify socially significant locations, and model organizational rhythms. Their work was based on a formulation for identifying structure in routine called *eigenbehaviors* [32]. By examining a weighted sum of an individual's eigenbehaviors, the researchers were able to predict behaviors with up to 79% accuracy. This approach also made it possible to calculate similarities between groups of individuals in terms of their everyday routines. With data collected in-the-wild over 100 days, Clarkson also presented an approach for the discovery and prediction of daily patterns from sensor signals [33].

While long-term activity prediction approaches have mostly relied on mobile phone data and sensor signals, our approach is focused on the prediction of human activities in real-wold setting from first-person egocentric images using computer vision and machine learning approaches. While there has been some work on detecting activities with egocentric cameras, most of these approaches rely on video and hand-crafted features. Fathi et al. [34] used egocentric video and detected hands and objects to recognize actions. Pirsiavash et al. [35] introduced an annotated dataset that includes 1 million frames of 10 hours

of video collected from 20 individuals performing activities of daily living in 20 different homes and used hand-crafted object detectors and spatial pyramids to classify activities using a SVM classifier.

In contrast to state-of-the-art approaches that use hand-crafted features with traditional classification approaches on egocentric images and videos, our approach is based on Convolutional Neural Networks (CNNs) combining image pixel data, contextual metadata (time) and global image features. Convolutional Neural Networks have recently been used with success on single image classification with a vast number of classes [14] and have been effective at learning hierarchies of features [36]. However, little work has been done on classifying activities on single images from an egocentric device over extended periods of time. This work aims to explore that area.

Privacy Concerns: One of the challenges of continuous and automatic capture of first person point-of-view images is that these images may, in some circumstances, pose a privacy concern. Privacy is an area that deserves special attention when dealing with wearable cameras, particularly in public settings. Kelly et al. proposed an ethical framework to formalize privacy protection when wearable cameras are used in health behavior research and beyond [37] while Thomaz et al. proposed a framework for understanding the balance between saliency and privacy when examining images, with a particular focus on photos taken with wearable cameras [38]. People's perceptions of wearable cameras are also very relevant. Nguyen et al. examined how individuals perceive and react to being recorded by a wearable camera in real-life situations [39], and Hoyle et al. studied how individuals manage privacy while capturing lifelong photos with wearable cameras [40].

2.3 Data Collection

Over a period of 26 weeks, we collected 40,103 egocentric images of activities of daily living for one subject with 19 different activity classes. The images were annotated manually using a tool we developed to facilitate this arduous daily task. The classes were generated

by the subject at their discretion based on what activities the user conducted (we did not provide the labels prior to data collection). The images were collected by recording a video at a rate of one frame every 60 seconds.

2.3.1 Process

The subject was equipped with a neck identity holder that was fitted to hold a smartphone in portrait mode. We developed an application that runs on the smartphone and captures photos at fixed intervals, which allows for the capture of egocentric data throughout the day. At the end of the day, the participant could filter through the images in order to remove unwanted and privacy sensitive images and annotate the remaining images. The participant categorized the data collected for 26 weeks using the annotation tool described in the following subsection into one of the 19 activity classes. The distribution of these classes is shown in Table 2.1. We can see that ”Working” and ”Family” are the top two dominant classes due to the participant's lifestyle. We note that the participant was free to collect and annotate data at their disclosure. The subject was also free to leave ambiguous images (i.e. going from work to a meeting) unannotated. Any unlabeled and deleted images were reasonably not included in the dataset.

2.3.2 Tool for Annotation

We developed a tool for rapid image annotation that is intended for daily activity labeling. The tool automatically receives the imagery taken from the application on the egocentric device and displays them in chronological order. The user is then able to select sequential images (in chunks) to label as specific activities. This facilitates the process of labeling large image sets in a simpler and intuitive manner.

Table 2.1: The distribution of the 19 different classes in the Daily Activities dataset.

Classes	**Number of Images**	**Percent of Dataset**
Chores	725	1.79
Driving	1031	2.54
Cooking	759	1.87
Exercising	502	1.24
Reading	1414	3.48
Presentation	848	2.09
Dogs	1149	2.83
Resting	106	0.26
Eating	4699	11.58
Working	13895	34.24
Chatting	113	0.28
TV	1584	3.90
Meeting	1312	3.23
Cleaning	642	1.59
Socializing	970	2.39
Shopping	606	1.49
Biking	696	1.71
Family	8267	20.37
Hygiene	1266	3.12

2.3.3 Description of Dataset

As shown in Table 2.1, the distribution of tasks is represented by a few common daily tasks followed by semi-frequent activities with fewer instances. We are keen to highlight the difficulty of certain classes due to their inherent overlap (socializing vs. chatting, chores vs. family, cleaning vs. cooking, etc). This class overlap is due to the inherent impossibility of describing a specific moment with one label (the participant could be eating and socializing).

The bi-weekly breakdown of data collection is shown in Table 2.2. We can see a general increase in the number of annotated samples later in the collection process. Some of this is due to increasing the interval at which the application captured images up to once a minute from once every five minutes. The rest of the increase can be attributed to the participant becoming more comfortable with the data collection and annotation process, and over time,

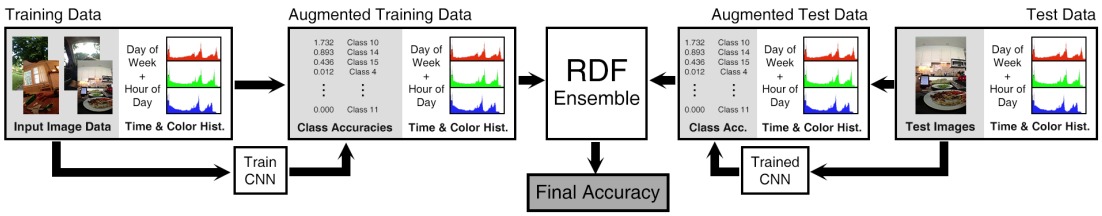

Figure 2.2: Overview of our Convolutional Neural Network Late Fusion Ensemble for predicting activities of daily living.

successfully incorporating this process into their day-to-day routine.

The participant collected the majority of the data from approximately 7am to 8pm. The majority of the data that is not captured is therefore during the participants sleep cycle. On an average day we retain 80% of the data that is collected (the participant removes approximately 20% for privacy and null classes). The participant handled null classes (blurry images, etc) by leaving them unlabeled. These images were then removed prior to assembling the dataset. It is important to note that the participant has still labeled images which we cannot release as they would violate the privacy of those involved (family, socializing, etc). For this reason we have opted for maintaining the dataset private.

Table 2.2: The bi-weekly distribution of the number of images in our Daily Activities dataset.

Classes	**Number of Samples**	**Percent of Dataset**
Week 1&2	553	1.40
Week 3&4	814	2.07
Week 5&6	69	0.18
Week 7&8	216	0.55
Week 9&10	239	0.61
Week 11&12	2586	6.58
Week 13&14	5858	14.90
Week 15&16	6268	15.94
Week 17&18	2903	7.38
Week 19&20	3417	8.69
Week 21&22	6465	16.45
Week 23&24	4695	11.94
Week 25&26	5229	13.30

2.4 Methodology

We present a methodology for incorporating contextual metadata and other traditional hand-crafted features with a Convolutional Neural Network (CNN) that processes the raw image data. The method is compared to baseline machine learning methods (k-Nearest Neighbors (kNN) [41] and Random Decision Forests (RDF) [42]) in order to demonstrate the benefits and limitations of our approach. We also introduce a method called late fusion ensembling for combining non-image data with CNN probabilities and compare it to a traditional CNN and classic ensembling methods.

2.4.1 Baseline Approaches

We ran evaluations using k-Nearest Neighbor (kNN) and Random Decision Forest (RDF) classifiers in order to adequately fine-tune the best accuracy for our baseline. We parameterized our dataset using contextual metadata (day of the week (as a nominal value from 0 to 6) and time of day) and global image features (color histograms). We found that a kNN classifier (with a k-value of 3) trained on the metadata and the color histograms (with 10 bins) gave an accuracy of 73.07% which was better than training a kNN trained on the metadata alone or the color histograms alone. We tested the classifier at incremental parameters of k (until 50) and found that performance slowly degraded as we increased k beyond 3. We further tested the time metadata at three granularities (the hour, hour + minutes (i.e. 7:30am = 7.5), and hour and minute as separate features) and found the difference in prediction accuracy to be negligible due to the scheduled nature of humans. We selected to keep the hour and minute as separate features as it had the highest accuracy. Further, we found that a RDF classifier with 500 trees trained on the metadata and color histograms (with 10 bins) gave us the best overall accuracy of 76.06% (note that random chance, by picking the highest prior probability, is 34.24% for this dataset). Training the RDF with more than 500 trees had a negligible effect on the total accuracy. Our baseline

Table 2.3: A comparison of the baselines using kNN and RDF trained on contextual meta-data (M), color histograms (H) and a combination of both.

	kNN M	**kNN H**	**kNN M+H**	**RDF M**	**RDF H**	**RDF M+H**
Avg. Class Accuracy	15.51	44.23	54.72	15.51	40.43	50.71
Total Accuracy	52.50	65.62	73.07	52.50	68.89	76.06

results can be seen in Table 2.3. It is important to note that a high total accuracy is driven by the distribution of the data amongst the classes. Since a majority of the data is in two classes ("Working" and "Family"), a classifier can achieve a high total accuracy by accurately classifying only those two classes. We also show average class accuracy to show how well the baseline classifier does for all classes distributed evenly.

2.4.2 Convolutional Neural Network

Recently, Convolutional Neural Networks (CNNs)[43] have been shown to be effective at modeling and understanding image content for classification of images into distinct, pre-trained classes. We used the Caffe CNN framework [44] to build our model since it has achieved good results in the past and has a large open-source community. Since the dataset has a small number of images, we fine-tune our CNN using the methodology of [45] using the ImageNet [46] classification model introduced by Krizhevsky et al. in [14] that was trained on over a million images in-the-wild. We retrain the last layer using our collected data with 19 labels for daily activity recognition. We set the base learning rate to 0.0001 in order to converge with our added data and use the same momentum of 0.9 and weight decay of 0.0005 as [14] with up to 100,000 iterations as shown in Figure 2.3. Our CNN has five convolutional layers, some max-pooling layers, and three fully-connected layers followed by dropout regularization and a softmax layer with an image size of 256x256 just as in [14]. We split our data by classes into 75% training, 5% validation, and 20% testing. The classifier was never trained with testing data on any of the experiments. The parameters were chosen using the validation set and the fine tuning in all of the experiments was only

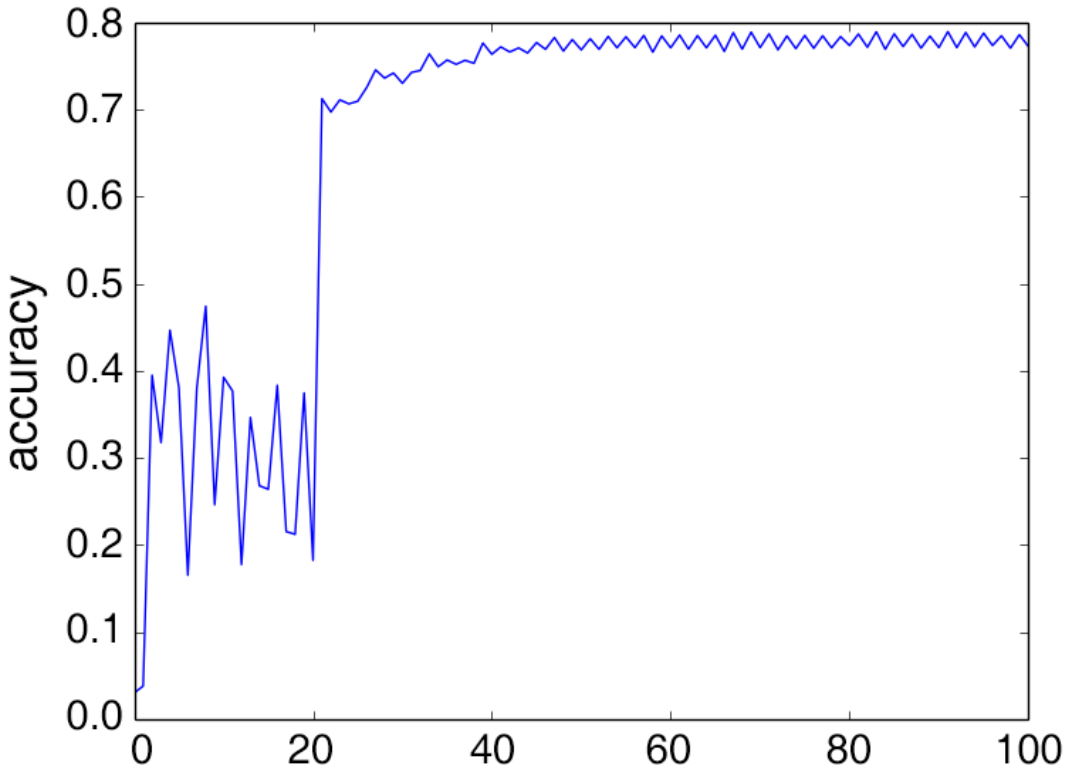

Figure 2.3: A Convolutional Neural Network trained for 100,000 iterations. We can see the accuracy convergence after 20,000 to 30,000 iterations.

done with the training set. It is interesting to note that the algorithm jumps to almost 78% accuracy after only 20,000 to 30,000 iterations and converges around 50,000 iterations due to fine tuning. Despite a high total accuracy, the class accuracy of a CNN alone is hindered due to the lack of contextual information and global image cues.

For many problems with small amounts of data, data augmentation can be effective at preventing overfitting and increasing accuracy. However, in this case, we are collecting data at a specific orientation and viewpoint, so many data collection techniques are not applicable. Because of this, we elected not to augment our training data although that would be a useful extension of the work.

2.4.3 Classic Ensemble

One method to combine the CNN output with non-image data is a classic ensemble method. Training a classifier such as a RDF on the contextual metadata can yield a probability distribution which can be combined with the CNN probability distribution to yield a final probability. This equally weights the CNN output and the RDF output in order to get the best output possible. This can prevent over-fitting from the CNN but doesn't necessarily increase the prediction accuracies since it doesn't leverage which classifier is better at which classes or which information from the classifiers are important.

2.4.4 Late Fusion Ensemble

To solve the problem of combining a CNN with a classic ensemble, we developed a late-ensemble technique. We use a RDF trained on the CNN soft-max probabilities along with the contextual metadata (day of week and time of day) and the global image information (histograms of color), each being separate features for the RDF. This allows for a good combination of outputs that can be learned rather than naively combined. Using this we outperform the classic ensemble and the normal CNN model by approximately 5%. The pipeline for our method is shown in Figure 2.2.

2.5 Results

In this section we present a comparison of baseline machine learning techniques against the different convolutional approaches for the classification of daily living activities. As shown in Table 2.3, kNN and RDF perform surprisingly well with contextual metadata (day of the week and time of day) and color histograms. RDFs marginally outperform the kNN methods, particularly with the use of color histograms. It is worth mentioning that we tested other global features (such as GIST [47]) on the same baseline methods and obtained negligible changes in accuracy.

Table 2.4: A comparison of different CNNs and CNN ensembles using contextual metadata (M), global features (color histograms (H)), raw image pixels and their combinations.

	Average Class Accuracy	**Total Accuracy**
Original	57.38	78.56
Ensemble (Pixel + M)	53.48	78.47
Ensemble (Pixel + M + H)	59.72	81.49
L-Fusion Ensemble (Pixel)	63.22	80.94
L-Fusion Ensemble (Pixel + M)	65.29	82.45
L-Fusion Ensemble (Pixel + M + H)	**65.87**	**83.07**

In order to improve the performance of our activity prediction we leverage the use of local image information. Through the use of a regular CNN, we see a minor increase in total accuracy (+2%) over the baseline (see Table 2.4), but a much more impressive jump in average class accuracy (+7%). We see an even greater increase in accuracy as we incorporate both contextual metadata and global image information (color histograms). We have demonstrated through the baseline methods that these features are of importance, which is why we developed our CNN late fusion ensemble that leverages the metadata and global and local image features. Our best ensemble leverages all of the presented information for a total accuracy of 83.07% with an average class accuracy of 65.87% showing an impressive increase over the baseline and the other methods. A confusion matrix of our final method's results is shown in Figure 2.4.

2.6 Discussion

Our method achieves the highest accuracy on the classes with the most samples (as one would expect since test accuracy increases with larger amounts of training data). As shown in Table 2.4, our ensemble method outperforms both a normal CNN and a classic ensemble with a CNN. Training an RDF with extra features and the CNN probabilities allows the RDF to find what is important for each individual class. It also allows for the other types of data to be effectively added in a framework that prevents some of the overfitting that

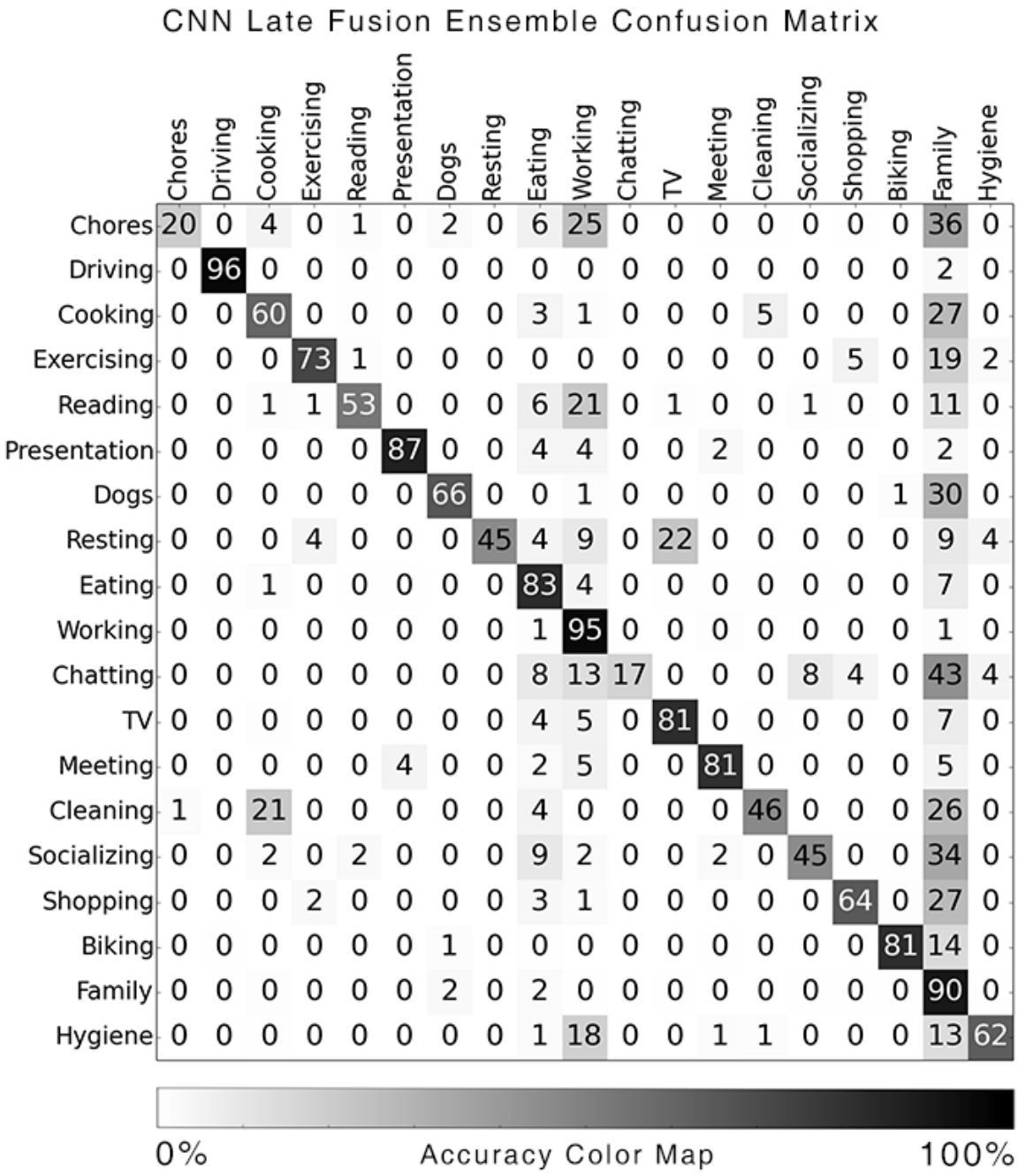

Figure 2.4: Confusion Matrix for the 19 classes of our dataset with columns as the predicted labels and rows as the actual labels.

CNNs typically have. This shows how our novel ensemble method effectively combines local pixel-level information, contextual information, and global image-level information. Because it relies on a CNN running on a GPU, the system uses a large amount of power and is not well suited for embedded devices. On an ARM device, testing each image would take

Table 2.5: A comparison of the best of all methods (using contextual metadata, color histograms and pixel data) for all the 19 activity classes. CNN+LF is CNN with Late Fusion Ensemble

	kNN	**RDF**	**CNN**	**CNN+LF**
Chores	**33.10**	17.24	00.69	20.00
Driving	55.07	60.87	**98.55**	96.62
Cooking	25.66	35.53	47.37	**60.53**
Exercising	44.00	63.00	69.00	**73.00**
Reading	**68.55**	49.12	30.04	53.36
Presentation	80.00	72.35	80.59	**87.06**
Dogs	62.17	44.35	55.65	**66.09**
Resting	**72.73**	54.55	27.27	45.45
Eating	77.14	75.75	82.05	**83.12**
Working	91.10	**96.42**	93.49	95.19
Chatting	**21.74**	04.35	00.00	17.39
TV	77.38	75.79	**81.75**	**81.75**
Meeting	68.73	61.00	73.36	**81.47**
Cleaning	26.56	30.47	38.28	**46.09**
Socializing	**52.85**	37.31	31.60	45.08
Shopping	40.16	27.87	63.93	**64.75**
Biking	19.57	23.19	78.26	**81.88**
Family	70.82	87.42	86.69	**90.15**
Hygiene	52.36	46.85	51.57	**62.60**
Avg. Class Accuracy	54.72	50.71	57.38	**65.87**
Total Accuracy	73.07	76.06	78.56	**83.07**

more than 15 seconds. However, the method could be run on a server that an embedded device could query.

Many of the classification failures of our method deal with some classes being interrelated. Our worst results are in “Chores” and “Chatting”. These classes can be easily confused with others such as “Cleaning”, “Working” and “Family”. In many examples in which the subject is conducting a chore, the family is in the background, which may confuse the classifier. We acknowledge this as a limitation of the method used for data capture that uses a single image frame in contrast to a short video clip. We believe the extension of our method to short video clips would prevent some of these difficult classification errors but would present further questions in privacy, device storage and battery life.

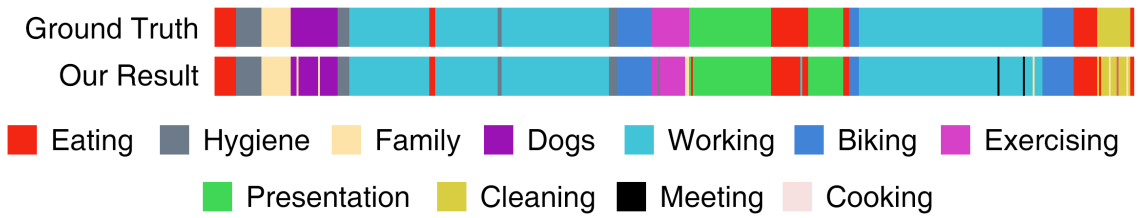

Figure 2.5: An example of a randomly chosen day and the classifier's predicted output.

However, there are specific motion features that differentiate certain classes from others. For instance, the reading and working classes are often done from a very stationary position where the camera is likely experiencing little to no movement. In contrast, biking and exercising encode a high level of motion in their frame due to the expected camera movement when performing these activities. These type of encoded features play a role in what the convolutional network is learning and how it categorizes specific actions.

To visually display an average day and our prediction of the activities for that day, we have taken a random daily sample from the data and classified it visually. The results are shown in Figure 2.5. For this particular day, our classifier nears 100% accuracy at predicting the user's activities. We are keen to highlight the misclassification errors on this given day. During the classification of "Dogs" at the beginning of the day (seen in purple), we notice two slivers of misclassification in which the algorithm detects "Family" instead of walking the dogs (both classes have instances of green foliage). We see similar errors in the last light blue segment, representing "Working", in which it detects two instances as "Meeting" instead of "Working". This provides further evidence that the class overlap is likely to contribute heavily to the 16.93% overall misclassification that we have in our dataset.

In a second experiment, we demonstrate the correlation between the amount of training data and the algorithms' test accuracy for the participant. We highlight two hypotheses for the increase in accuracy over time. The first is that the algorithm is adequately learning the participants' schedule and frequented activities, which allows it to better model their daily

activities. The second plausible hypothesis is that the algorithm is adapting to general human behavior and learning the overall characteristics of specific classes. This presents two interesting questions for the applications of this research. First, how much data is required to train a generic model and second, how much data is required to "fine-tune" said generic model to a specific user. We have tried to address the first of these questions by training our model with varying amounts of data points to observe the number of days/samples a user is required to collect in order to train a good generic model. The top 7 classes are shown in Figure 2.6 (plots for the other 12 classes are omitted to maintain clarity). We can see that the class accuracies improve as more data is captured with a significant increase in accuracy after the first 4 weeks.

In order to address the second question, we performed a final experiment in which two volunteers (V1 and V2) wore the egocentric device for 48 hours in order to collect images and time-stamps at a 60 second interval. The data was divided equally into a training and test set (Day 1 for training and Day 2 for testing) in order to test the validity of the model trained by our original participant's data. The results of this experiment are demonstrated in Table 2.6. As you can see, for some classes that involve a similar viewpoint and environment, like reading, the model generalizes very well. However, for many others such as driving and chatting where volunteers are going different places and talking to different people, the model does not generalize well. It is worth noting that the initial accuracy prior to fine-tuning performs worse than the highest prior probability of the original model (34.24%). We reason that this is due to the difference in habits between participants (we work, read, cook for distinct periods of time) that require fine-tuning to adapt to one's specific daily schedule.

Different individuals also have different activities and one set of class labels from one individual might not fit another individual's lifestyle. Given the model trained for one person we wanted to study the possibility of fine-tuning a classifier to yield good results for a different person that performed non-overlapping activities. At its core, this addresses

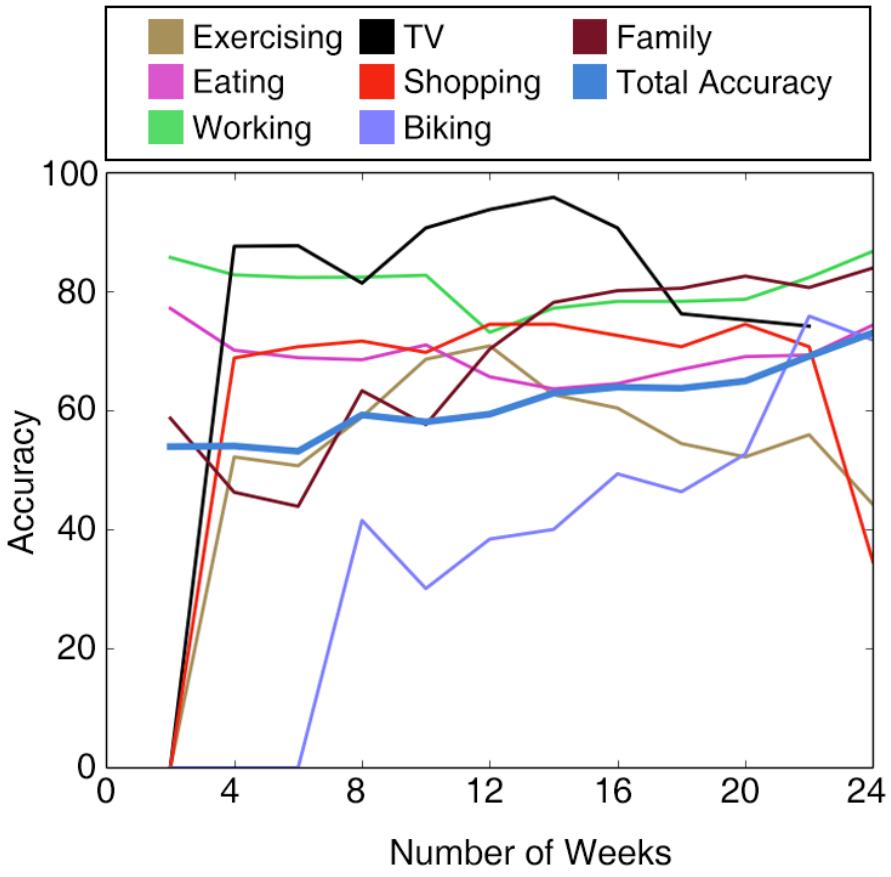

Figure 2.6: A plot of class accuracies vs. the number of weeks of training samples. We can see a general trend where the class accuracies increase as the amount of training samples increase. A significant increase in accuracy is seen after training on the first 4 weeks of data.

the question of whether a classifier is learning the schedule and habits of one person or if the learning is inherently adapting to common human behavior. As seen in Table 2.6, the classifier trained on the original participant was not very successful. However, fine-tuning that model with just one day of data from the new user can yield very good accuracy. Not only did this achieve great accuracy, but the CNN converged in less than 5,000 iterations, whereas the original CNN took more than 50,000 iterations to converge. This implies that part of the model is learning human behavior while another part is learning the habits of a specific person. We can use a small amount of training data to fine-tune the classifier to learn the habits of a new person, while still keeping the knowledge of general human

Table 2.6: A comparison of the original model tested on two volunteers and the fine tuned model. "Original" is the original applicants data and model. "V1" and "V2" are the results from the original model tested on volunteers 1 and 2 data respectively. "V1 Fine" and "V2 Fine" are the results from the fine-tuned models trained on volunteers 1 and 2 data respectively. The results that are not available are classes that the two volunteers did not perform when collecting their data.

	Original	**V1**	**V1 Fine**	**V2**	**V2 Fine**
Chores	20.00	5.56	25.0	N/A	N/A
Driving	96.62	18.6	100.0	0.0	100.0
Cooking	60.53	0.0	25.0	N/A	N/A
Exercising	73.00	0.0	50.0	N/A	N/A
Reading	53.36	77.78	75.0	N/A	N/A
Presentation	87.06	N/A	N/A	N/A	N/A
Dogs	66.09	N/A	N/A	N/A	N/A
Resting	45.45	N/A	N/A	N/A	N/A
Eating	83.12	11.48	76.92	30.68	100.0
Working	95.19	31.59	98.32	39.14	94.44
Chatting	17.39	0.0	86.67	0.0	96.72
TV	81.75	0.0	33.33	N/A	N/A
Meeting	81.47	0.0	100.0	0.0	60.0
Cleaning	46.09	0.0	0.0	N/A	N/A
Socializing	45.08	0.0	0.0	0.0	83.33
Shopping	64.75	40.0	50.0	N/A	N/A
Biking	81.88	N/A	N/A	N/A	N/A
Walking	N/A	0.0	57.14	N/A	N/A
Family	90.15	N/A	N/A	N/A	N/A
Hygiene	62.60	13.33	0.0	27.78	81.82
Class Acc	65.87	10.56	51.83	13.94	88.05
Total Acc	83.07	23.58	86.76	27.06	91.23

behavior.

2.7 Conclusion

In this section of our work, we have demonstrated a robust and unique dataset of egocentric images that have been annotated with the user's activities, a CNN late-fusion ensemble method to classify image data with relevant contextual information, promising results in fine-tuning the model to other users and a trained model that performs well on egocentric daily living imagery. We have shown state-of-the-art results on the data compared to

commonly-used methods (a traditional CNN and a Classic Ensemble) and we have determined the amount of data that is needed to train an initial CNN classifier for this problem and the amount of data that is required to fine-tune the model on a per-user basis. Given that our model was able to fine-tune with such a little amount of data demonstrates that it is not simply encoding visual parameters that are unique to an individual but also encoding high-level features, such as motion, that are representative of the activity. This lead us to further research the implicit and explicit role of motion in video classification.

CHAPTER 3
MOTION IN DANCE

This section of our work was published to the arXiv library on January, 22, 2018 [17].

3.1 Introduction

Video is a rich medium with dynamic information that can be used to determine, what is happening in a scene. In this work, we consider *highly dynamic video*, video that requires the parameterization of motion over extended sequences to identify the activity being performed. The main challenge with highly dynamic video is that a single frame cannot provide sufficient information to understand the action being performed. Therefore, multiple frames, leading to an extended sequence of frames, need to be analyzed for scene classification. One of the drawbacks of current action classification research is both a lack of approaches that can be applied to extended/long sequences and datasets lacking in such highly dynamic videos. Our goal is to determine which methods best represent motion as opposed to methods that use a single (properly picked) frame [48] to identify the activity, as we feel such approaches devalue the necessity for video data. In this work we introduce a 1,000 video dataset and evaluate methods that focuses on highly dynamic videos requiring motion analysis for classification.

We choose the domain of dance videos as (a) there is large amount of dance videos available online and (b) the diversity of dynamics in these videos provides us with a challenging space of problems for highly dynamic video analysis. This enables us to conduct a focused study on the relevance of motion in dancing classification and the broader value of motion in improving video classification. The core challenge of this task is attaining an adequate representation of human motion across a 10-second clip. In order to highlight the trajectory of this work, we will evaluate the current approaches and demonstrate the value

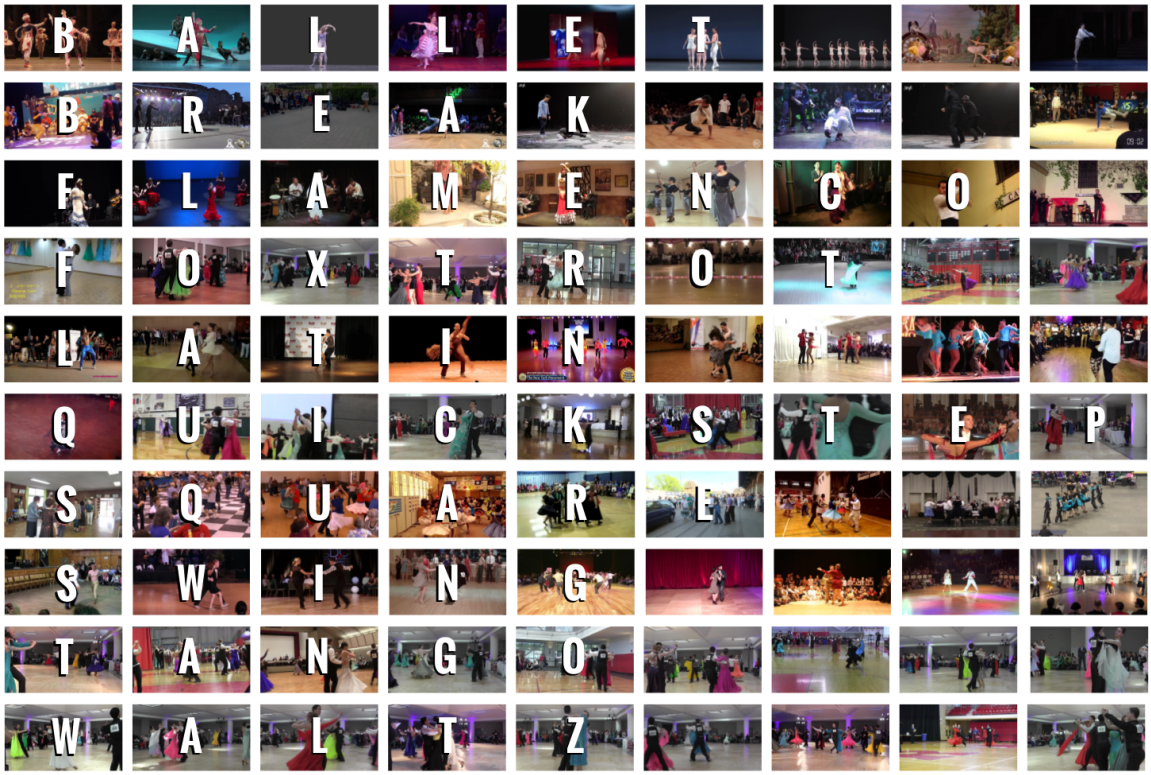

Figure 3.1: Each row contains frames from the class it represents. This figure is best viewed digitally.

of isolating motion for properly evaluating these approaches and this dataset.

Many video classification techniques exist, either utilizing single frames, late fusion architectures, temporal (3D) convolutional networks, or recurrent networks with long short-term memory (LSTM). Current classification problems can often be identified by a single frame. We present a more challenging problem wherein each class requires the use of multiple frames to adequately classify each category. Specifically, we propose the use of optical flow and multi-agent pose estimation as motion representations which augment traditional video classification approaches. Comparing these approaches enables us to gain insights into the inherent encoding of motion in neural networks that is difficult to understand.

Our main contributions in this chapter are: **(1)** An analysis of baseline and state-of-the-art approaches in video classification, **(2)** a general method for concurrently learning from multiple motion parameterizations in video, and **(3)** A 1000 video dataset of highly

dynamic dance videos, contrasted with existing video datasets, to motivate further investigation and understanding of motion parameterization in video classification.

3.2 Related Work

In order to determine which competing state-of-the-art approaches to examine, we first present a literature review on video classification. While deep networks have been shown to be very effective at classifying, localizing, and segmenting images, it is still unclear how to properly extend these methods to the video domain. There are a few common approaches, some of which are: (1) Applying proven image classification deep network architectures to individual frames of a video; (2) Extending 2D convolutional operators to 3D convolutions acting on the time domain; and (3) Preprocessing the video into images that encode motion, such as optical flow, and running current image architectures on the processed frames.

A simple way to extend image-based neural networks to video classification is to extract features from each individual frame of a video [49]. While this technique does lead to some success if the network learns temporally-invariant features, it is commonly only used as a baseline approach to compare against networks that incorporate temporal data [50, 51]. One common variant is a two-stream late fusion architecture with a still frame-based "spatial" network stream running in parallel alongside a "temporal" network performing classifications based on optical flow calculations [52, 48, 53, 54]. This network architecture significantly outperforms approaches based solely on individual frame classification, suggesting that incorporating a temporal component is necessary. In our work we leverage the benefit of a temporal network by incorporating it into the design of our network architecture.

Karpathy et al. explore more direct methods of incorporating temporal data with each video frame by extending the convolution kernels from size m x m x 3 to m x m x 3 x T, where T represents a temporal extent[51]. They also point out one of the major challenges

of using deep learning for video classification – there are no large-scale video datasets comparable to the quality and size of image recognition datasets. Similarly, 3D convolutional kernels that incorporate the spatial domain have been shown to be successful for action classification in both security camera and depth data recordings [50, 55]. Wang et al. use a similar two-stream late fusion approach [56], but they note that without incorporating the learned features into an ensemble method with handcrafted features, these deep-learned approaches still fail to outperform handcrafted approaches. We combine these methods in our work by incorporating preprocessed features (optical flow and multi-agent pose detection) with 3D convolutional kernels in order to integrate the representation of motion into the network architecture.

Another common approach is to leverage the sequential nature of a Long Short-Term Memory (LSTM) network–a specific type of recurrent neural network with additional gates to control the flow of information. LSTMs can process information over long term temporal sequences and have been applied in video for various tasks such as caption generation [57] and learning video representations [58]. Similarly, the long-term recurrent convolutional networks (LRCNs) proposed by Donahue et. al. introduce another variation of an LSTM for this task. Despite their temporal nature, these approaches have been less successful at encoding motion in comparison to two-stream networks [59] which encode the spatial and temporal domain in concurrent architectures.

The most effective method for classifying motion in video is still unclear. In the context of action recognition, many of these approaches are learning features based on the image's context and not the inherent action. This is in part because commonly used video datasets such as UCF-Sports and more traditionally UCF-101 can generally be identified to moderately decent accuracy using single-frame approaches which do not encode motion parameters[51].

A specific method for encoding motion that has recently gained traction in action recognition is the use of pose detection over the temporal domain with neural networks

Figure 3.2: Each of these examples represents a different class in our dataset (they are types of ballroom dancing). Top Left: Waltz, Top Right: Quickstep, Bottom Left: Foxtrot, Bottom Right: Tango.

[52][60][61][62]. Detecting pose over this domain provides us with the intrinsic motion of the subjects in the scene. As highlighted earlier, an initial breakthrough was achieved by Toshev et. al. [61] with state-of-the-art results in estimating the pose of a single individual from an image. The importance of pose was further demonstrated in [52], incorporating pose features from a CNN into action recognition. This work was extended over the next two years to attain joint-specific networks that work well with partial and occluded poses [62]. It was then most recently implemented to detect multiple people within a single frame [63]. In our work we will leverage our own implementation of multi-agent pose detection to demonstrate the need for motion parameterization when classifying highly dynamic video.

3.2.1 Existing Datasets

There are a handful of relevant datasets that exist in the research domain. We highlight some of the more relevant video datasets that are appropriate to our work. All of these

datasets demonstrate the growing need for understanding what type of motion features are relevant in classifying highly dynamic actions, which we explore in this work.

UCF-101

The UCF-101 dataset [64] contains approximately 13,000 clips and 101 action classes, totaling 27 hours of data. The clip length varies largely from 1 second to 71 seconds depending on the activity at a resolution of 320x240. This was one of the first datasets to tackle human actions in video. However, as we will demonstrate in this work, most per-frame (image-based) approaches still perform moderately well on the dataset, illustrating the main question which we seek to answer in this work – that being the representation of motion as a classification feature.

Kinetics

The Kinetics dataset [65] contains 300,000 clips and 400 action classes, with a minimum of 400 videos per class. The action classes are also loosely grouped in 32 parent classes which further break down the dataset. This dataset was collected semi-automatically with curation through image classifiers and use of Amazon Mechanical Turk to determine the action classes and if the video snippet was appropriate to that class.

Atomic Visual Actions (AVA)

The AVA dataset [66] contains 80 atomic visual actions in 57,400 movie clips which are localized within the frame. This work goes beyond simply understanding a simple action in a video clip to understanding the interaction, both between humans and with humans and objects. Although this is somewhat less relevant to our work, it demonstrates the need for understanding motion features in human interaction – specifically by localizing the action and its relevance in a scene that may contain multiple subjects / objects.

3.3 Let's Dance Dataset

Our main challenge in this work was determining a reliable way of testing how well a specific method can parameterize motion. We realized that available video datasets such as UCF-101 [64] and UCF-Sports [67] tackled a known classification problem, one that could be evaluated using extensions of available image classification architectures.

With that in mind, we developed a new dataset that prioritizes motion as the key characteristic of the classification. We assembled a 1,000 video dataset containing 10 dynamic and visually overlapping dances. We chose the parent category of dancing because it has a variety of measurable features (rhythm, limb movement), and it is not represented in the Sports-1M and UCF-101 datasets [51, 64]. The categories included in this dataset are:

- Ballet
- Break Dancing
- Flamenco
- Foxtrot
- Latin
- Quickstep
- Square
- Swing
- Tango
- Waltz

The dataset contains 100 videos for each class. Each video is 10-seconds long at 30 frames per second. The videos themselves were taken from YouTube at a minimum quality of 720p, and includes both dancing performances and plain-clothes practicing. Examples of each class can be seen in Figure 3.1.

We highlight that the dataset contains four different types of ballroom dancing (quickstep, foxtrot, waltz, and tango) as seen in Figure 3.2. The motivation behind picking these dances is that their parent category is specifically the setting in which the dance occurs (a ballroom). This satisfies our main challenge of selecting classes that exemplify highly dynamic video. Quickstep is a very fast-paced type of ballroom dancing that is considered upbeat. Foxtrot is a much more fluid and continuous type of ballroom dancing, performed with a $4-4$ time signature. Waltz is quite similar in style, but it is performed with a $3-4$

time signature, leading to an increasingly difficult problem to tackle. Tango is yet another type of ballroom dance that originates from South America due to European immigration. Some of it is set to have originated from Waltz, but took on a different style over the years with a mix of a variety of other dance types. On this note, we extract two different motion representations from our input data for use by the community; optical flow[68] and multi-person pose detection.

When attempting to detect pose, we found numerous methods that focused on single-person pose detection. We adapted these methods to multiple individuals (given that dancing is generally a group activity, see Figure 3.3) through the use of a recent real-time person detector[69]. Similar approaches can be seen in [70][71][63].

After detecting the bounding boxes for each person in the scene we computed the pose for each individual using [62]. Positive and negative examples of this methodology can be seen in Figure 3.7.

3.4 Baseline Methods

In order to better understand the need for motion parametrization in video, we have identified two commonly-used architectures to establish as our baseline. These are architectures which are commonly applied to video architectures but only take a single-frame as input (per architecture).

These approaches are extensions of very successful image classification techniques.

3.4.1 Frame-by-Frame

Using the architecture of a state-of-the-art convolutional neural network for image classification, such as VGG [72], a classification for the video can be achieved based on key image frames from a video. A sample architecture based on CaffeNet, a variation of AlexNet [73], is shown in Figure 3.4. This approach does not explicitly encode motion in determining the video's classification but rather categorizes each frame and naively selects the majority

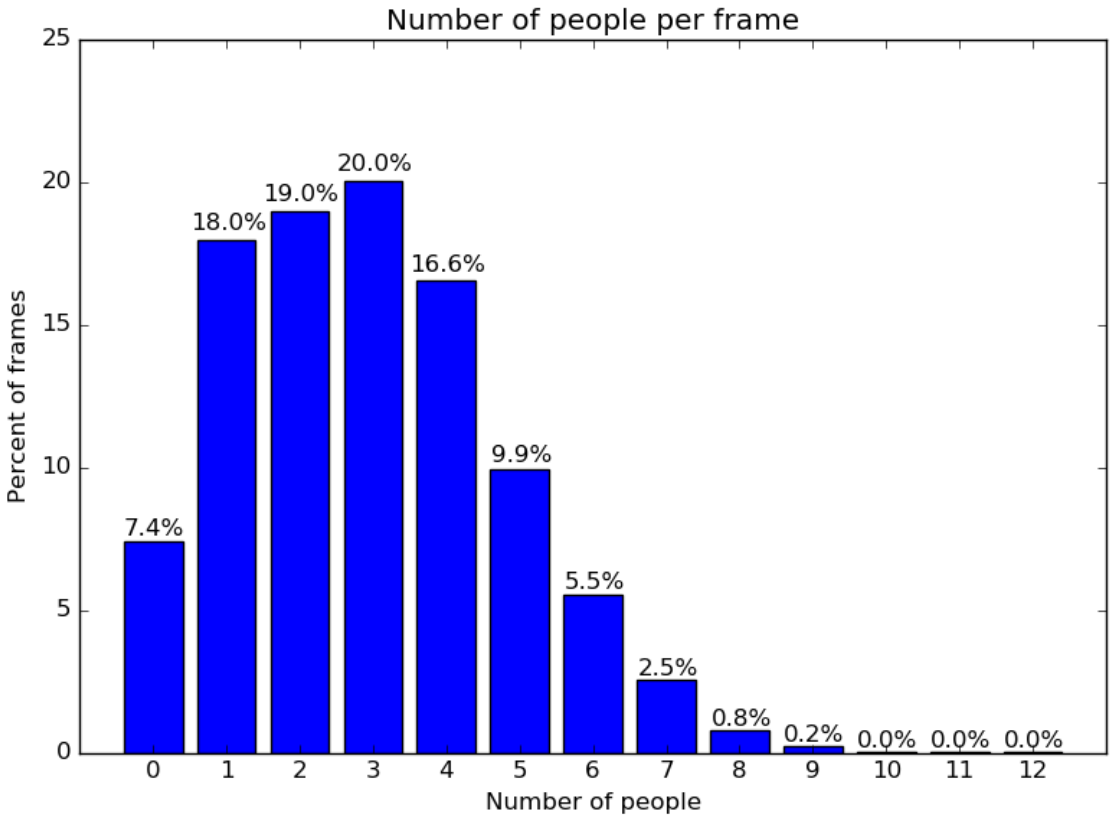

Figure 3.3: Distribution of number of people per frame using [69]. 75% of frames had at least two people detected in the dataset. 56% of the dataset has more than two people in the shot, which further illustrates the added complexity of this task.

label.

We do note that although there are numerous approaches for aggregating a single class from multiple per-frame classifications, the network itself does not encode the temporal domain.

3.4.2 Two-Stream Late Fusion

A common way of adding a temporal component to deep networks is by separately performing a classification based on spatial data (a single frame) and temporal data (i.e. optical flow). Merging these results produces an overall classification for the video, as shown in Figure 3.5.

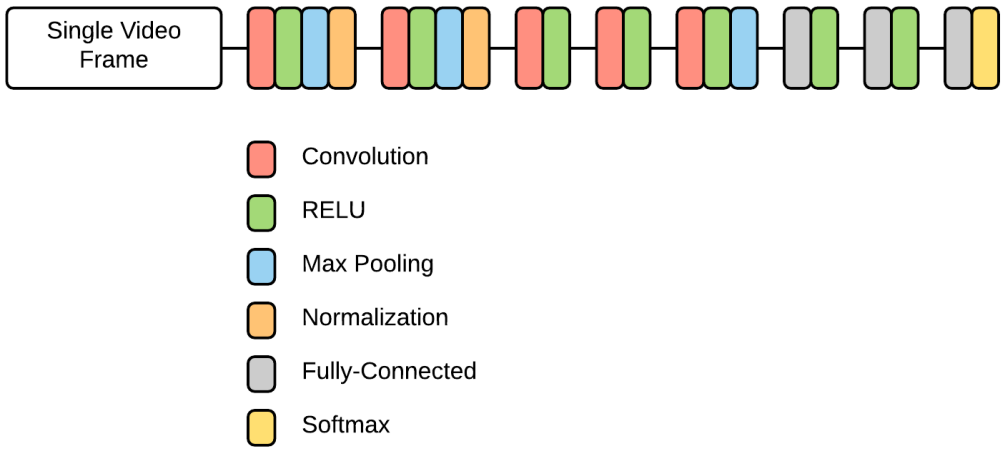

Figure 3.4: Frame-by-Frame Architecture: This is a traditional CNN, commonly used in image recognition.

This approach computes optical flow from two frames (at time n and $n - k$ where k is not necessarily 1) over the period of the entire video. Each frame in this case can be considered a single instance of motion that occurred in the video. For dancing we envision this as a specific move in a dance.

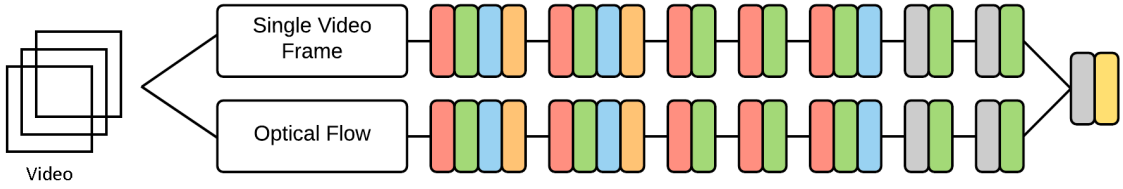

Figure 3.5: Two-Stream Late Fusion Architecture (color key in 3.4). This method incorporates motion (optical flow) into a traditional CNN pipeline.

3.5 Proposed Approaches

In order to address the challenge of categorizing highly dynamic videos we implement a number of methods which explicitly encode motion. At the core of these approaches is the notion of 3-dimensional kernels which process a series of video frames for classification. This enables us to pass in very short video clips (16 frames or approx. 1/2 second) for the network to learn. The overall objective was to incorporate motion in the learning pipeline

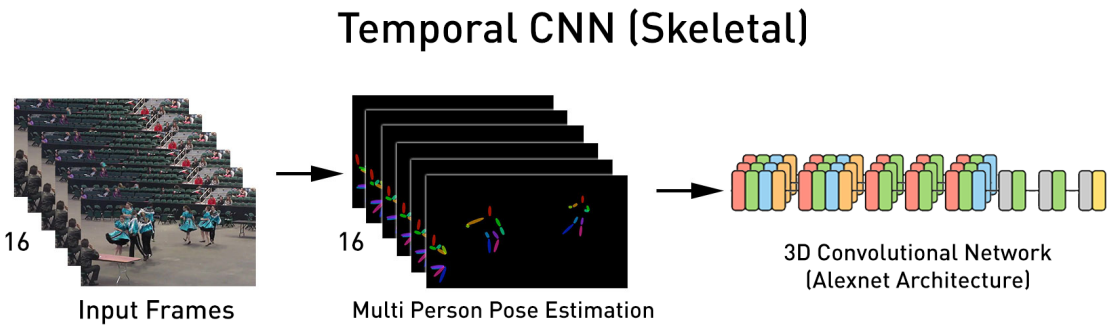

Figure 3.6: This pipeline displays a skeletal temporal CNN (3D Convolution) which processes the initial frames to obtain a multi person pose estimation from the input frames obtained by performing a bounding box person detection from [69] which is then processed by [62] for detecting the dancers' pose.

of standard approaches and assess their performance. After testing these approaches it was evident that combining numerous motion parameterizations in a concurrent deep network architecture would best represent the input video.

3.5.1 Temporal 3D CNN (RGB)

As stated, traditional convolutional neural networks can be extended to video by using 3-dimensional kernels that convolve in the temporal domain. We focus on testing this slow-fusion approach discussed in [50], which embeds the high-level spatial and temporal information at the initial convolutional layers by propagating the information through the network. One of the main setbacks of this proposed approach is the computational time it currently takes to compute these methodologies. We discuss this further in Section 3.7.

3.5.2 Temporal 3D CNN (Skeletal)

In this pipeline we compute a temporal CNN on multi-person pose information. We visualize the pipeline in Figure 3.6. This architecture demonstrates the importance of leveraging context for particular videos. Dance videos inherently benefit from this representation given that there are generally multiple people in the scene. Through the use of a visual-

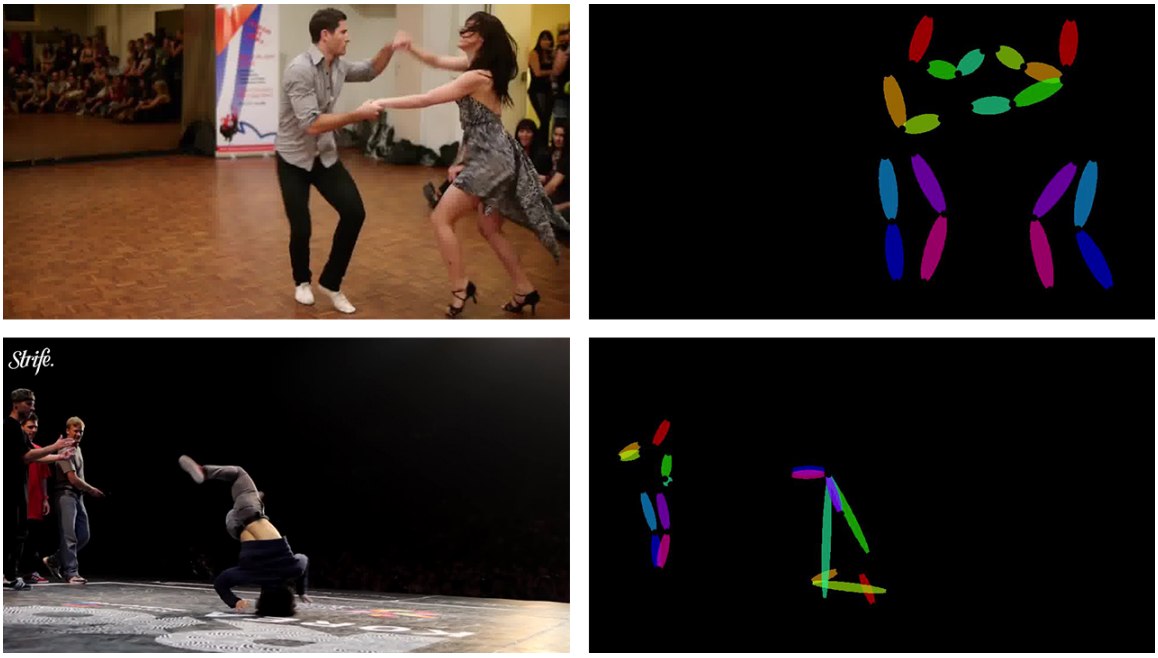

Figure 3.7: Demonstration of outputs from our pose detection pipeline. Top: Latin dancing positively classified. Bottom: Break dancing being erroneously classified. The dancers' left leg is accurate but his remaining limbs fail.

ization of pose we are able to attain comparable results to single-frame CNN approaches. It is key to note that this method does not use visual information from the original frame but solely visualized pose information as shown in Figure 3.7. Similar to our optical flow approach, it is likely that this method benefits heavily from encoding the number of people in the frame in addition to the motion over the 16 frames that are convolved in the temporal domain.

3.5.3 Three-Stream CNNs

We tested both single-frame and temporal approaches for a three-steam convolutional network in order to directly compare the potential importance of embedding multiple frames into the learning pipeline in addition to providing multiple representations of your original input. We highlight that these temporal convolutions are computing 2D convolutions over each of the input frames. Although this increases the complexity of our model it still remains significantly more tractable than computing 3D convolutions which require

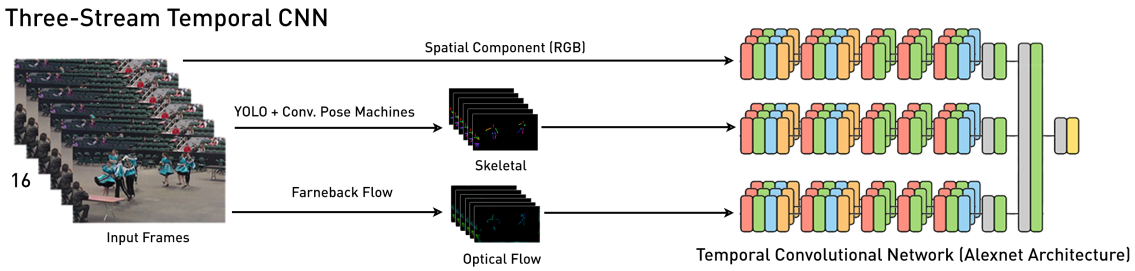

Figure 3.8: This visualizes the workflow for our three-stream temporal CNN which uses three convolutional stacks to process the spatial and respective motion components. It aggregates the fc7 layers into one and outputs the dance classification for a 16 frame input.

approximately twice the computational power.

Frame-by-Frame Architecture

For the frame-by-frame architecture, the first stack of our network processes the spatial representation of our input which is our RGB image. Our second stack processes the optical flow representation which was computed from frames n and $n - 10$ in order to accentuate particular motions from a given dance. Our third stack processes our multi-person pose visualization explained in Figure 3.6. As discussed earlier, this stack is essentially encoding the number of participants detected for a given dance frame and their current pose.

Temporal Architecture

The temporal architecture utilizes the same three stacks but processes chunks of 16 frames at a time in order to incorporate a temporal component into the loss of the network. This enables us to learn motion parameters from the spatial, optical flow and multi-person pose representations. A visualization of our pipeline can be seen in Figure 3.8 whose convolutional and fully connected layers are based on the standard AlexNet architecture[73].

3.6 Baseline Experiments

We implement our proposed approaches with the goal of determining which approach is most effective at highly dynamic video classification. Implementation details for each approach are given below.

3.6.1 Dataset Splits

We extract individual frames from the Let's Dance dataset (1000 10-second videos at 30fps, resulting in 300000 frames), which we then randomly split per video into 80% training, 10% testing and 10% validation (consistent across experiments). Optical flow and pose detection was split in the same manner in order to consistently test the approaches.

3.6.2 Frame-by-Frame

To perform a baseline video classification experiment, we implemented the architecture shown in Figure 3.4 in Tensorflow[74]. The weights for the network's convolutional layers are initialized to values from a network pre-trained on the ILSVRC 2012 dataset [75]. Final video classification results can be determined by classifying each frame in a video and voting to determine the video's overall class.

For an initial comparison, we also tested the network with optical flow imagery as the input.

Overall, we observed significant amounts of overfitting in the original training accuracy which hints at the network learning too much about the appearance of the specific videos in the training set for each class. As we hypothesized, using image frames alone results in the network learning features that do not generalize well to the dancing categories, since it has no way to observe the motion inherent in the video. Testing accuracy peaks at 56.4% over 10,000 iterations of fine-tuning the network. We compare these results to a similar framework introduced by [76] which tested the frame-by-frame baseline on UCF-101, attaining

an accuracy of 72.8%. This directly demonstrates the possibility of solving the classification problem by carefully selecting the right frame versus understanding the underlying motion of the video.

We also ran the identical setup using optical flow estimation. Before training we precompute optical flow for the entire dataset. We used Farneback's method for calculating dense optical flow [68] to obtain a per-pixel estimate of the horizontal and vertical components of motion and then incorporate this into the same network architecture.

In this case we saw slightly worse performance at approximately 45% for testing. We do note that the overfitting for optical flow images is subdued given that the per-frame images no longer contain background information. Given that a number of our dances occurred in similar or identical settings, background information was a strong confounding factor for the original images. The overall result for optical flow performs worse than training on RGB images given that it is merely embedding the motion between two frames. We will later demonstrate that larger frame chunks provide significant improvements to this approach.

3.6.3 Two-Stream Late Fusion

We implemented the two-stream late fusion architecture shown in Figure 3.5 in Caffe[44]. The two-stream approach follows intuitively from the previous subsection in which we discuss the effects of both a frame-by-frame method on images and on optical flow. Each individual stream uses the CaffeNet architecture, with weights initialized to a network pretrained on the ILSVRC 2012 dataset [75]. We then fine-tune the network by training only the fully-connected layers at the end of each stream, which are then concatenated and passed through a final fully-connected layer which outputs the respective classifications.

We note that each architecture in the two-stream method is still using a single frame as input, and as such the network is trained on a frame-by-frame basis. We chose to use the CaffeNet architecture for each frame, initialized with the ILSVRC 2012 weights, to be

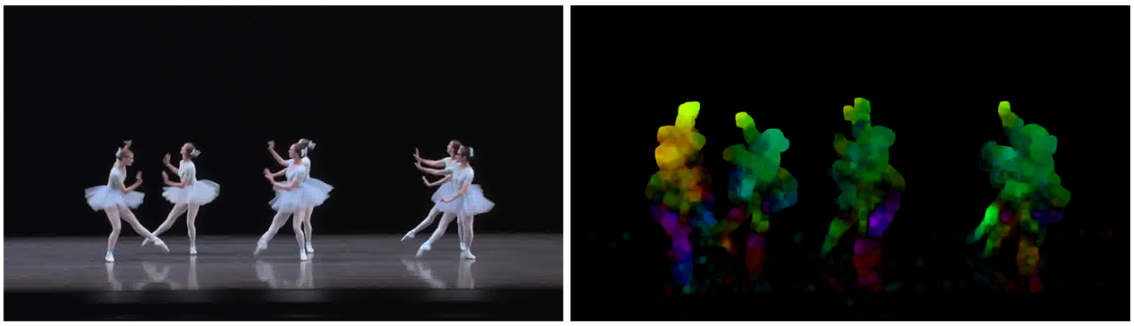

Figure 3.9: An image of dancers performing ballet and their optical flow estimation. As we can see, optical flow does a good job of segmenting the subjects in the scene in addition to encoding their motion.

consistent with the baseline frame-by-frame experiment described in the previous section. This allows us to perform a direct comparison between the two-stream and frame-by-frame approaches, to determine the benefit of optical flow on this dataset. As with the frame-by-frame approach, final video classifications can be determined by classifying each individual frame and optical flow image pair, followed by voting to determine an overall class. It is interesting to note that the total per video classification accuracy of this method was 68.89% which is much higher than the single frame-by-frame accuracy of 56.40%. Although one may be compelled to argue that single-frame motion is key to this classification, we refer back to Figure 3.3. This figure demonstrates that the frames throughout the dataset also contain a tremendously varied number of participants. As we can further see in Figure 3.9, optical flow tends to visually separate the dancers from the background, which also explains the significant increase in the algorithm's performance. In addition to the motion in a single frame pair, the foreground's shape and representation is playing a key role in the classification of the network.

The results demonstrate an improvement over each independent approach, with a classification accuracy of 64.69% per-frame. This is a significant increase of 10% above the imaging method and 20% over the optical flow method. This increase was attained by combining the same architecture as the previous two methods, with the addition of a single concatenation node to fuse the data at the end of the network. It demonstrates that

directly incorporating temporal data into a network can be immediately beneficial towards classifying video.

Leveraging the network to perform full video classification (rather than only per-frame classification), we tested the trained network on our test set of videos, taking the class with the largest number of per-frame votes as the final video label. This resulted in a per-video classification accuracy of 66.14%. After further experimentation with the network architecture, we saw a significant improvement from computing a unique mean image to subtract from the optical flow, which increased our accuracy to a final per-video classification rate of 68.89%.

The network performs well at classifying Ballet, Waltz, Tango, Flamenco, and Foxtrot, with poor classification accuracy on Break and Swing dancing. Of particular interest is the network's performance on Waltz, Tango and Foxtrot which occur in similar settings. As such, the network shows that it's capable of performing fine-grained classification within the Let's Dance dataset.

Table 3.1: Method Comparison of UCF-101 and Let's Dance. UCF Frame-by-Frame results obtained from [76], Two-Stream results obtained from [54]

Dataset	**Frame-by-Frame**	**Two-Stream**
UCF-101 [64]	**72.8%**	**88.0%**
Let's Dance	56.4%	68.89%

Lastly, we revisit our accuracy results with UCF-101, a well-established activity recognition dataset. Table 3.1 illustrates high levels of accuracy on UCF-101 using the standard extensions of image classification techniques which we discuss in this section. It is important to note that the two-stream comparison is comparing a two-stream accuracy for UCF-101 that utilizes an SVM to combine its streams whereas we concatenate the final layers of both convolutional streams into the fully connected output. As stated earlier, this illustrates the core issue we encountered in looking for a highly dynamic dataset which further validates our motivation to introduce the "Lets Dance" dataset to the research com-

munity.

3.7 Results & Discussion

In order to assess our temporal architectures we compare with a number of state-of-the-art approaches that explicitly encode motion in order to determine their performance. Overall it has become clear to us that we need to transition from traditional per-frame CNN approaches when conducting video classification.

It is evident from Table 3.2 that methods which embed motion significantly outperform traditional methods and that metrics to evaluate these approaches are necessary in order to better understand what each network architecture is learning.

3.7.1 Temporal 3D CNN

In order to evaluate this approach we restructured our data into 16-frame chunks that were needed as the input for the 3D convolution. The network could be trained on the 3D features from 16-frame non-overlapping chunks of the video. We fine-tuned from the network trained on UCF101 by [77]. This method yielded a per-video accuracy of 70.11%. This result was particularly impressive because it demonstrated the inherent ability of a 3D convolution to extract motion features that are not explicitly computed. The major drawback of this approach is its complexity. A 3D convolution inherently takes significant computation for a single-stream.

We were unable to perform multi-stream approaches using 3D convolutions due to this complexity. In order to combat this we introduce more tractable approaches for state-of-the-art graphics cards (Our current systems utilizes Titan Z Pascal graphics cards) that achieve comparable performance by explicitly encoding motion into the network architecture. In addition to this we note that 3D convolutions are limited to the initial input-size which in our case was 16 frames. This makes it difficult to encode more complex motions that last more than 1/2 second without sub sampling frames which will invariably lead to a

Table 3.2: Comparison of numerous approaches and their testing accuracies on our dataset

Approach	**Testing Accuracy**
Frame-by Frame CNN	56.4%
Two-Stream CNN	68.89%
Temporal 3D CNN (RGB)	70.11%
Temporal 3D CNN (Skeletal)	57.14%
Three-Stream CNN	69.20%
Temporal Three-Stream CNN	**71.60**%

loss in detail. Most temporal methods will invariably suffer from this limitation given that variable inputs into a convolutional network has not been fully explored.

3.7.2 Skeletal Temporal 3D CNN

In order to embed human motion data, we incorporate skeletal images into a temporal CNN. We visualize each pose into a single image which represents the pose for that particular frame. We attained an accuracy of 57.14%. We note that this accuracy still performs marginally better than a frame-by-frame approach despite the fact that it does not utilize the spatial (RGB) representation. Due to the computational complexity of running concurrent 3D convolutional networks we propose a stacked 2D convolutional method which allows us to combine multiple streams in a single state-of-the-art graphics card.

3.7.3 Frame-by-Frame Three-Stream CNN

Our Three-Stream Frame-by-Frame architecture utilizes all three data modalities. We assess this as both a single-frame and as a stacked architecture in order to compare their benefits and drawbacks. As shown in Table 3.2, this approach attains an accuracy of 69.20%. This three-stream network performs comparably to the two-stream fusion approach we conducted as one of our baselines which indicates that there is not a significant amount of information added from the use of both skeletal and optical flow representations.

3.7.4 Temporal Three-Stream CNN

Logically, we extended our frame-by-frame approach into the temporal domain by stacking the image input layers to produce a 16-frame chunk. This approach utilizes the same input as the Temporal 3D CNN we implemented at a much lower complexity for three streams. We saw this method attain the best performance out of all of the methods we evaluated, at an accuracy of 71.60%.

Looking at our most successful approaches, three-stream methods and 3D convolution, we note that both achieve very similar performance in per-video classification. However, the two methods are not equivalent in terms of computational resources. Beyond the increased workload and restrictions inherent in appropriately formatting the data for the temporal CNN, 3D convolution is much more computationally-intensive at both training and testing time. We observe that even though the temporal CNN was our most successful approach, it may be sub-optimal when a much simpler three-stream stacked convolutional network approach is available.

3.8 Summary

In this work we sought out to understand the effect of motion on classifying videos. Recent work in the are has demonstrated the relevance of these type of videos, most recently seen in [66] and [65]. The work we have conducted demonstrates that traditional CNN approaches do not properly or intentionally encode motion in their methodology. This fact is frequently overlooked by testing on videos that do not inherently require motion. That was the primary motivator of this work. As we can see in Table 3.2, 3D convolution methods outperform more traditional approaches by inherently encoding motion into their computation and prediction. Similarly, two-stream methods that incorporate optical flow can also leverage temporal features to significantly improve video classification.

This also opens up some potential for future work in incorporating optical flow and

pose data. Hybrid approaches, such as a three-stream temporal CNN, have the potential to increase an algorithm's understanding of the video. We have also developed a more focused dataset that we believe the research community will benefit from by intentionally selecting highly dynamic actions in one specific class. We tested a variety of traditional and more complex methods in order to begin to understand the composition of our dataset and its baseline performance. The Let's Dance dataset will continue to help us to assess adequate motion parameterization and hopefully assist in improving how we learn from video data.

One of the biggest problems we ran into throughout this research endeavor was determining the best classes to select for our dataset. Initially we had some intuition for dancing and martial arts being adequate parent categories but we quickly saw that martial arts represented a multi-class problem. Although dancing exhibits similar overlaps the separation was much more evident when performing the data collection. We also had to alternate between different dances partly due to availability on YouTube and our own understanding of these dances.

In the next chapter we focus on explicitly learning from the motion of a human pose. We develop a parameterization of a human pose and incorporate this parameterization alongside existing deep neural network models by leveraging our previous work on late-fusion ensembles in egocentric images. This technique demonstrates the improvements that our parameterization has on existing learning approaches. We also increment the difficulty and complexity of our dataset to further highlight the value of motion in video classification.

CHAPTER 4

LET'S KEEP DANCING: PARAMETERIZING POSES IN HUMAN ACTIONS

4.1 Introduction

The previous sections in this thesis cover the recognition of human actions from two different settings. The first setting was a single participant conducting daily activities with an egocentric camera. We were able to accomplish this classification by leveraging contextual information about the participants coupled with the image data using a technique known as late fusion. The second setting was from a third-person perspective in which the subject(s) were dancing and the task was to recognize the specific action (type of dance) they were engaging in. In this chapter, we were able to generate two additional modalities derived from the video which contributed to the understanding of the action, optical flow and human poses.

Although optical flow demonstrated some improvement in understanding the action, there was little intuition beyond a high-level parameterization of motion. Human poses, obtained by leveraging state of the art work in detecting human poses from video [63], were a trickier challenge because it was unclear how to integrate into classification. A naive attempt used a pose visualization in a standard neural network to learn on a high-level how the dancers moved. Here we discuss the additional work on leveraging pose information in a logical manner that enables one to specify specific joint arrangements that are relevant to the action and in our case, to dancing. In addition to this, we further utilize our prior work on daily activities to couple this information with image data to achieve significant improvements in recognizing unique dances. We test combining the various different modalities using standard fusion and late fusion approaches, and present our findings to the research community.

Human pose estimation is the process through which we estimate a set of vertices that represent joints of the human body from a digital image. This is an intuitive representation of a human being as it mimics our skeletal structure that does not explicitly encode particularly relevant information for learning about the action of the subject. For instance, in dancing, the position of your arms and legs and how it changes over time is part of what determines the type of dance you are performing. This posture and its change over time is part of what we are interested in when studying these actions. Therefore, developing features that encode these motions accurately is the core objective of our work.

In previous work we have explored determining these actions by incorporating relevant information to the task. When looking at egocentric images in order to determine your daily activities, we incorporated the time of day and day of week as incredibly relevant features that determine what activity you may be engaged in (most people are at work on Tuesday at 10am) [18]. We utilized these features alongside a neural network that learned specific patterns from the actual image using a late fusion ensemble which significantly improved the performance of our algorithm. Following that work, we developed the Let's Dance dataset [17] where we collected highly-dynamic videos to test the parameterization of motion in video. This chapter brings together the implicit encoding of motion and contextual parameters from Chapter 2 with the work we did in testing existing deep neural architectures on a highly dynamic dataset which we introduced in Chapter 3.

4.2 Related Work

The ability to recognize a human being from a series of joints was first popularized by Johansson et al [12] in the 1970s. In this work, Johansson proposed extracting joints from a walking subject by shining a flashlight at reflective tape in a darkened environment. In a controlled setting, Johansson tracked the joints of a subject moving in a single horizontal plane from the left to the right of the image frame. By treating these points as moving vectors, he was able to demonstrate that these simple dots over time did in fact encode a vast

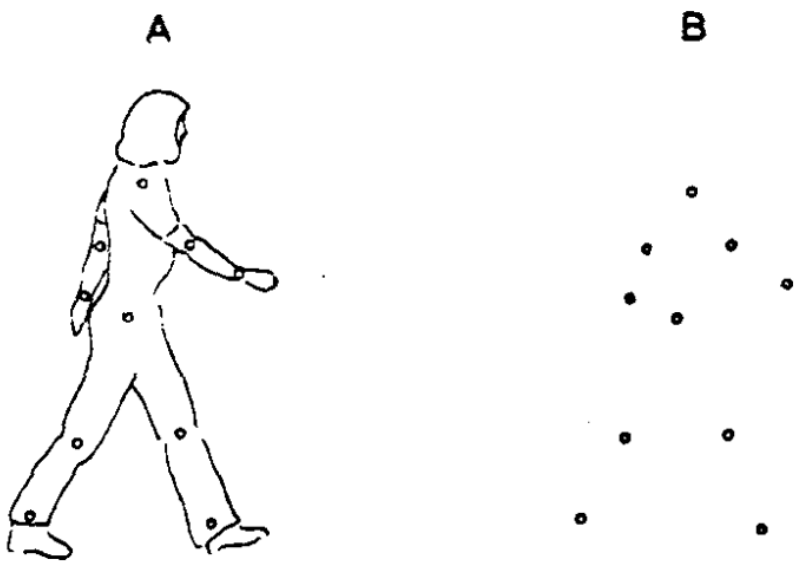

Figure 4.1: Source: [12]. This figure represents one of the first attempts at representing a human pose using joints in 2D space.

amount of information about an individual's motion and be indicative of the activity which they were performing. Figure 4.1 demonstrates a drawing of his setup which contained a total of 10 joints. These joints are still the basis of how we compute a human pose today. The work of Shotton et. al. using depth [78] was a pivotal moment in the vision community with the introduction of a consumer-friendly depth sensor which could detect a human pose with incredible levels of accuracy. Additionally, the work of Toshev and Szegedy using neural networks on digital images [61] further improved our detection of a human pose from an RGB camera. Today, pose detectors can accurately detect a large number of human poses in a single frame in real-time. In this work we will explore the benefit of leveraging these representations to improve our understanding of human actions.

Prior work in the field of action recognition has demonstrated the overall importance of high-level features (joints) in improving our understanding of specific actions [79] and therefore the need to develop appropriate representations. One of the prevailing issues in this area is the evaluation using datasets with a high level of inter-class variability. For

instance in [80] they evaluate their method using the Penn Action dataset which compares baseball, golf, tennis, and specific workouts (very distinct poses / classes). Similar issues were prevalent in the KTH and UCF datasets [81][64] which contained fairly simple unique actions. We do note in this work that the UCF dataset contained some classes which overlapped in their setting, but most of them differed significantly. Although these datasets have been studied at length [82] [76], it was clear that more complexity was required to be applicable to the real world. This issue was further highlighted by Rohrbach et al [83], who looked at fine grained activities in cooking. In this work, a series of body model features are introduced which improve the overall understanding of specific actions. We seek to extend some of these features and combine them with neural network techniques in order to improve our overall understanding of motion in a scene.

4.2.1 Existing Datasets

There have also been high-level efforts to introduce challenging action recognition datasets in this field that are in a similar realm to our work. The Kinetics dataset [65] took a hierarchical classification of their data by providing approximately 600+ classes (this is continuously growing) and has kept the number of examples to be at least equal with the number of classes (so at least 600 videos per class). These classes are grouped into parent classes to produce a hierarchical mapping for each of their unique classes. One of the main issues with the dataset is that a lot of the data is not as precise in encompassing each action. The dataset was collected semi-automatically and used raters on Amazon Mechanical Turk to assess the appropriateness of the video snippet to a category.

The AVA dataset [66] is another effort which focuses more on human interaction. This dataset defines atomic actions with respect to the interaction between humans or a human and an object. This scratches well into the surface of the problem of understanding complex activities and begins to understand the relationship between objects in a scene, an area which we find of tremendous value.

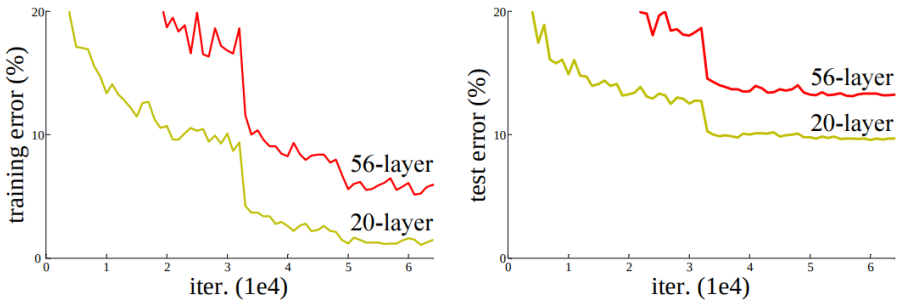

Figure 4.2: Source: [84]. This figure, originally presented by He et. al., demonstrates the issues with simply applying complexity can backfire as networks stop learning as effectively.

In between both of these datasets lay a niche area which we felt encompassed our dataset well, and that was the exploration and understanding of human motion for these activities. Therefore, in the first part of our work we introduced the Let's Dance dataset [17] as an effort to increase the difficulty of detection between unique human poses and to introduce complex multi-person actions into the domain. We tested the different modalities of our dataset (rgb, depth and pose) using standard deep learning techniques in order to establish a baseline for our dataset. This was one of the objectives of this dissertation. The next objective was to improve our classification by improving how we parameterize a human pose in video and how we combine that with RGB techniques.

4.2.2 Existing Deep Learning Approaches

In the ever-growing space of deep learning, there are numerous approaches to understand and classify images. On a high-level, deep learning involves layers of mathematical operations that attempt to determine patterns on an matrix of values. This matrix of values can be a representation of captured light (an image) but it is not limited to those measures. It could also be a computation of depth (via optical flow, LIDAR sensors, etc) or a visualization of a human pose. Given the vast number of modalities that can be represented in a matrix of values, there are a massive number of approaches in the research space. It is important

to note that our objective was not to reinvent this metaphorical wheel or develop a new specialized layer that incremented accuracy by a certain percentage. Our approach focused strictly on the proper understanding of our methods and the use of intuition to improve the accuracy of existing neural network models. We will briefly overview some of the most common approaches, some of their benefits and drawbacks, and a high-level reasoning as to why we selected a particular approach.

Convolutional Neural Networks

Krizhevsky et. al. popularized the use of neural network in their seminal work on ImageNet classification in the early 2010s [14] [46]. Like most seminal work, it was built on decades of research in the area. One of the first instances of convolutional networks dates back to the work of Fukushima [85] which wisely describes how these networks are able to learn on their own. A few years after that, Yann LeCun applied similar methods combined with backpropagation to tackle the task ok handwritten digit recognition [86]. Although initial network structures such as AlexNet were comprised of 5 convolutional layers, a lot of the following innovation relied on the ability to make these networks deeper and maintain the networks ability to properly learn without overfitting. In 2015, Szegedy et. al. followed up on their state-of-the-art work on the Inception networks (also known as GoogLeNet for Inception v1 [87]) and published Inception v3 which leveraged factorization to decrease the number of parameters in the network whilst maintaining its complexity [88]. By this point, the network was 42-convolutions deep and had introduced batch normalization and factorization (which also made the network wider). Today, it is not uncommon to see networks with over 100 convolutional layers in depth, and additional novelties to prevent the challenges that come with an incredibly deep network.

Residual Networks

One of those changes is residual connections. In 2016, He et. al. [84] leveraged residual networks to improve the predictions of very deep networks. As image datasets continued to expand in complexity and size, researchers continued attempting to make networks deeper and deeper in an effort to learn more fine-grained detail. This general approach of 'just going deeper' eventually fell apart as very deep layers were no longer learning additional information. Depth mapped fairly well to the size of a dataset but not directly to the performance. Common issues arose after tackling a problem with excessive complexity. Networks commonly over-fitted due to the number of variables used as shown in Figure 4.2. It was clear that deep networks were having trouble learning from layer to layer given that they appeared to be at maximal depth for the problem at hand (this problem is often referred to as the degradation problem). We will revisit this concept when training networks for the Let's Keep Dancing dataset, where readily available 100+ layer networks quickly over-fit the problem due to what we theorize is a sheer excess of parameters and a lack of more data (from the perspective of our dataset in comparison to a dataset the size of ImageNet). Residual networks essentially combine the output of convolutional layers in depth, meaning that they mathematically aggregate the weights at certain stages in the network (applying pooling or reshaping operations as necessary to match in size) to better propagate information forward. The result enables network to learn better information when facing the degradation problem. We address this network in our work because although it is commonly used, we found that for smaller datasets it seemed to encounter excessive problems in successfully training.

4.3 Let's Keep Dancing: Expanding the Let's Dance Dataset

One of the problems we chose to tackle in the pose detection and action recognition field was the high inter-class variability and the actual necessity for video data. Many of the

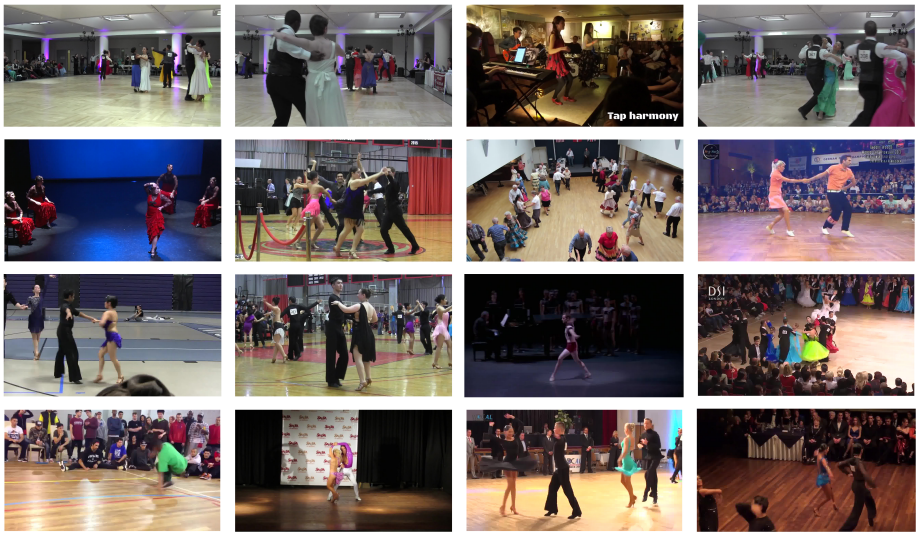

Figure 4.3: The following figure represents a snapshot of each of the 16 dances presented in our most recent work. From left to right, then top to bottom: Foxtrot, Tango, Tap, Waltz, Flamenco, Samba, Square, Swing, Cha, Rumba, Ballet, Quickstep, Break, Latin, Jive, Paso Doble

initial classification datasets did not require motion to classify their categories (i.e. you could argue that classifying baseball vs soccer can be achieved using a single image, it is certainly the case that humans are able to make that distinction). We provided a solution to this gap in datasets by manually collecting 1600 videos which represent the Let's Dance dataset. This dataset, which originally contained 10 types of dances, now contains 16 dances with added styles of ballroom dances to increase the difficulty of the problem. The main issue with categorizing these classes is that visually it is difficult to distinguish the subtle difference between each dance. The classification therefore relies in understanding the motion of each of the dancers over time. As we can see in Figure 4.3, many of these dances occur in plain clothes, and are often difficult to identify on their own. Due to the lack of domain knowledge we were unable to test how well humans would be able to identify each individual dance, but we did test their interpretation of motion for two pose representations and found that identifying the parent activity was highly likely.

Table 4.1: The distribution of the 16 different classes in the Let's Dance dataset. The discrepancy in the number of videos is due to video attrition suffered on the YouTube platform over the last 3-4 years of collecting and maintaining the dataset. If a user takes down their video we respect the extent of that request and remove it from our dataset.

Classes	**Number of Videos**
Ballet	89
Break	95
Cha	98
Flamenco	88
Foxtrot	79
Jive	106
Latin	90
Pasodoble	98
Quickstep	82
Rumba	94
Samba	97
Square	97
Swing	95
Tango	80
Tap	95
Waltz	80

4.3.1 Optical Flow Estimation

In the original Let's Dance dataset, we were using Farneback's approach for optical flow estimation [68]. After introducing six additional dances, we opted for recomputing optical flow for the entire dataset with a more modern approach to obtain more high quality flow information for our analysis. In this update, we used FlowNet 2.0 as an improvement over the original optical flow estimation, example results of which can be seen in Figure 4.4. This improvement in flow has the added advantage of improving how well subjects are segmented from the background in incredible detail. As we can see in the first row, the center figure is segmented to pixel-level accuracy in much higher quality than standard approaches to detect flow.

Figure 4.4: Left: Original Frame. Middle: Farneback's Optical Flow Estimation [68]. Right: FlowNet 2.0 Output [89]. The smoother segmentation of subjects is shown on the first row with a more challenging example demonstrated in the last row where the algorithm struggles with a darker scene.

4.3.2 Pose Estimation

Similarly, we improved our computation of human poses to a more modern approach that was able to more consistently detect a human pose. Our original approach to pose detection relied on first detecting a bounding box for each person using YOLO [69] and then processing that image with a convolutional network to detect each joint [62]. We improve this pipeline with a more recent approach from Facebook Research called DensePose [90] that seemed better equipped for complicated scenes. As we can see in Figure 4.5, the detections are more prominent and better visualized to identify a unique person in the scene.

Visual Consistency of Poses

These visualizations are generated by extracting the 2D joints on a per-frame basis from each clip. After extracting the poses, we assign each pose a color based on the minimal

Figure 4.5: Left: Original Frame. Middle: Pose Detection from [62]. Right: FlowNet 2.0 Output [89]. The smoother segmentation of subjects is shown on the first row with a more challenging example demonstrated in the last row where the algorithm struggles with a darker scene.

pixel distance between the current and previous poses.

$$color(p_i) = \begin{cases} \min\limits_{j=i-1,\ldots,j=i-k} d(p_i, p_j), & \text{if } d(p_i, p_j) < thresh \\ \text{new color}, & \text{otherwise} \end{cases} \tag{4.1}$$

Here, we compute a thresholded distance metric between the current pose and the poses detected in the previous k frames. For each pose, we compute their distance to the previous poses and if they are within the distance threshold we assign that skeleton the color of the previous frame. We can experimentally determine the distance threshold by making an assumption about how quickly a human can move in k frames. It is important to note that frame rate is not always consistent in the dance dataset so this metric should be computed per video if a specific distance wants to be enforced. The variable k can be determined experimentally by visually assessing the persistence of the pose detection algorithm from frame to frame. We determined $k = 2$ for DensePose [90]. A threshold of 10 pixels worked well for our dataset. This approach has certain drawbacks in the context of dancing. When

couples are dancing, we see issues both with detection and misassigned colors when one dancer is identified as the other. This happens most commonly when the pose detection algorithm fails to detect both dancers correctly but does not consistently detect one or the other. We fine tune our network on these visualized frames, which are consistent with the frames that we presented to participants in Experiment 2.

4.3.3 Specificity of Classes

One of the issues that has come up since the release and analysis of our dataset is how specific a particular label is in the context of dancing. We would like to acknowledge a mistake we made in labeling a dance type as latin', which doesn't do a good job of encompassing a particular type of dance but rather a high-level category. About 50% of this category is in fact bachata, whilst the remaining half is salsa. In retrospect, we should have split these categories up. Additionally, we are keen to acknowledge that many of these dances have similar origins and likely overlap in tempo, movement, and style. We see this as part of why it is such a difficult and compelling problem to tackle.

4.4 Human Experiments: Mechanical Turk Studies

In order to better understand how human beings extracted information from human poses, we conducted two experiments. The purpose of the first experiment was to understand whether we could extract intent from moving dots. This was a modern replication of a study that was done decades ago and is heavily referenced in this work by Johansson et. al. [12]. The second experiment was to determine whether more accurate classification could be extracted from our pose visualization. In this visualization it is more clear that the subjects are human, and the goal is more targeted towards determining the specific activity they are undergoing.

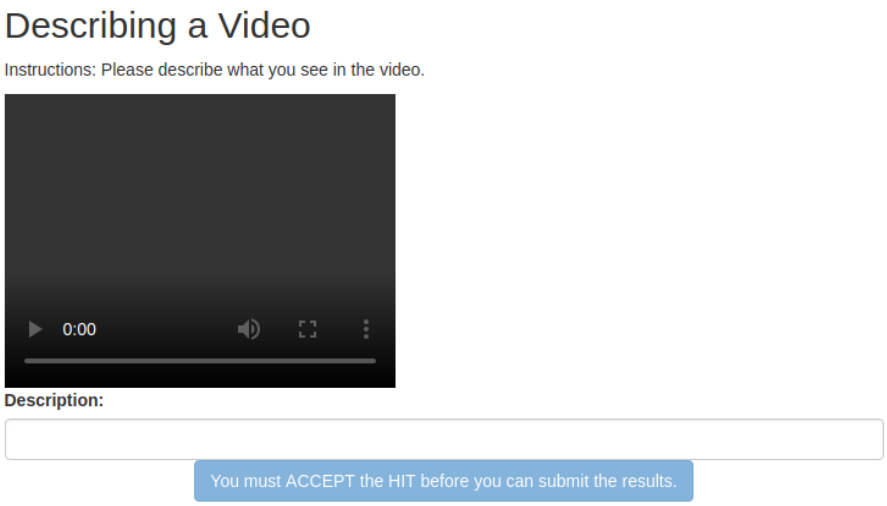

Figure 4.6: Prompt shown to Turkers with a short clip of moving yellow dots, as visualized in Figure 4.7.

4.4.1 Experiment 1: Understanding Simple Poses

During our first Mechanical Turk experiment we explored whether humans could identify a person and the activity they were conducting based on a visualization of their joints over time. The main objective is to understand if the motion of a series of pixels encodes a persons' activity. We were surprised that every participant was able to identify the moving dots as a human being and also classify the activity they were performing accurately. Some of the responses are shown below in table 4.2, with all of the responses included in our Appendix.

In this experiment, we asked participants to describe what they saw in the video (prompt shown in Figure 4.6). It is important to note that we made no reference to human beings, or activity recognition, we simply asked for a description of what they saw. We chose the Weizmann Human Action dataset [91] which contains 10 simple actions in a fairly controlled environment that would be straightforward to detect (bending down, jumping jacks, jumping, hopping on one foot, running, moving sideways, skipping, walking, and waving with one and two hands). After running the original video through a simple pose detection

Figure 4.7: Left: Original frame from the Weizmann Human Action dataset, subject is waving with both hands. Right: A frame from the video shown to Mechanical Turk participants, depicts the human pose.

algorithm, we visualized the joints as yellow dots and asked participants to describe what they saw in the video. We obtained a total of 279 responses from 19 participants. Despite no indication of the content of the video, an overwhelming majority of participants were able to label the action correctly and all of the participants assumed the dots were a human. Humans were even able to extract intention from seemingly basic classes. For instance, for the action of a participant waving both their hands, shown in Figure 4.7, a subject responded with " Man said bye or don't come here". It is fascinating to see that we are not only prone to assume that it is a person, but also their gender and what they are potentially communicating. The main objective of this high-level study was to confirm that we are not only capable of determining that these dots are of a human being, but also the action that the participants are conducting. We opted for a simple action recognition dataset and not our own dance dataset for this study because it was likely participants would be unable to pick out which dance they were observing without prior knowledge but providing that knowledge would bias them to know that it was a human action.

Table 4.2: Some examples of Mechanical Turk responses from our first experiment. As you can see, people are always accurately describing the motion of the figure. The only arguably incorrect responses were people describing running as walking, or subtle differences like describing skipping as hopping.

Original Video Class	**Turker Response**
Walking	The figure is casually walking.
Skip	The figure is hopping on one leg.
Jump	Jumping on both feet
Galloping Sideways	A figure is side-stepping to the right.
Two-hand Wave	One man say bye or don't come here.
Skip	A person hopping on their right leg and moving forward
Jumping Jacks	a man doing jumping jacks with the left arm jerking a bit unnaturally
Running	The figure is running to the left of the screen.
Jumping in Place	Man just jumping in one place

4.4.2 Experiment 2: Can we see human poses dance?

The second experiment we conducted visualized the human poses using DensePose [90], a more recent technique for detecting a human pose. In this experiment we asked participants to describe the action being performed in the video as specifically as possible. Although we do not expect most participants to identify the specific type of dancing, given how the first experiment went, we did expect users to categorize all activities as a type of dancing, or simply dancing. The idea behind this experiment was to understand if human's could determine that the motion of a human pose was correlated to dancing. The audio for each of the videos was not included in order to control for the bias that would likely introduce to this experiment. The setup was nearly identical to Figure 4.6, with the only difference being that we added a flicker warning for any participant that may have been sensitive to flickering colors (visually, a pose would often flicker when a detection was missed in a particular frame).

As we can see in Table 4.3 (see Appendix for full list of responses), most participants recognized that the individuals were dancing, and some even named types of dances they thought were particularly similar to what they saw. It was impressive to see people pick

Table 4.3: Some handpicked examples of Mechanical Turk responses from our second experiment. Responses left as is. These examples were picked to demonstrate some of the best descriptions for the pose videos we got.

Original Video Class	**Turker Response**
Foxtrot	figures ballroom dancing
Flamenco	This is dancing stage performance and the guys dancing near the guy using music instrument.
Quickstep	A solo dancer dancing in a semi circle followed along by other solo dancers. all the dancers are moving their hips like if dancing to salsa.
Break Dancing	this video clip single guys only dancing like a hip hop and other are watching.
Tap	All guys are dancing in the floor.this is group dance.beautiful co-ordination
Ballet	this video clip 3 members only dancing like that bharatham dance in he floor
Break Dancing	A solo dancer in the middle moving his feet really fast. A few dancers in a line also moving their feet fast. A solodancer break-dancing their way to the middle

out break dancing and ballroom dancing from the list of classes. This table is however hand picked as good results, the majority of results mention dancing in some capacity but do not go into as much detail about the type of dance being performed. These responses demonstrated that human beings are able to determine that the poses are dancing, but likely due to a lack of domain expertise were unable to determine the specific type of dance. These results helped us understand the knowledge that is embedded in the motion of a human pose and motivated our work in parameterizing it for improving video classification.

4.5 Methodology

We present a standard methodology for the classification of these 16 dance categories. We tested a number of existing approaches and present a late-fusion approach to combine our motion parameterization to demonstrate improvements on a standard network. In comparison to previous work on this dataset, we make use of significantly deeper networks and leverage the improvements in our data representations to obtain the best performance for

Figure 4.8: Left: Original RGB Image. Center: Estimated Optical Flow [89]. Right: Human Pose Estimation [90]

the Let's Dance dataset. In this section we will overview that approach and present our results.

4.5.1 Data Representations

The data was processed into three representations: original RGB frame, optical flow representation obtained via FlowNet v2.0 [89], and a visualization of poses extracted from DensePose[90]. An example of these three representations can be seen in Figure 4.8.

4.5.2 Baseline Approach

For each representation, we fine tune the Inception v3 neural network model [88]. This model is pre-trained on ImageNet labels (1000 classes) and we fine tune it to our dataset over 4000 iterations. We resize all of our input images to the same size as the model input, which is 299 x 299 in resolution. For all experiments we divide our data into 80% for training, 10% for testing and 10% for validation. We are always keen to note that we train and test on a per video basis, meaning that videos that were seen in the training set are not in the testing or validation set. We divide the sets equally per class so they contain roughly the same number of samples per class for every stage of the learning process.

4.5.3 Fusion Approaches

For our results we leverage two types of fusion to combine different modalities. For standard fusion approaches, we use the same network and simply combine the network predictions to determine the classification, which is how networks are normally combined to incorporate additional information. For late fusion approaches, we use a random decision forest to combine network predictions together. A late fusion approach tends to demonstrate improvements over standard fusion approaches because it enables the learning algorithm to decide which prediction may be more valuable for certain classes, which is key in improving performance.

Three-Stream + Parameterization Late Fusion Approach

Our final network uses three Inception-v3 networks trained on ImageNet and finetuned for our RGB data, optical flow, and dense pose. We use the class predictions of each of these networks as features into a random forest which we combine with the parameterization of the pose shown below. We show that this late fusion approach enables a straightforward way of embedding intuitive features into neural network approaches to increment our understanding of what the network is learning and how we can improve performance with more informed intuition. The network is shown on Figure 4.9.

4.5.4 Parameterization of a Human Pose

From here, we developed a set of parameters that would properly represent the joints we were looking at. The main goal was to develop a representation that would uniquely identify a human pose. The list of joints that is commonly detected in pose estimation is the following:

- Head (more recent algorithms detect nose, ears, and eyes).
- Shoulder (Right and Left)

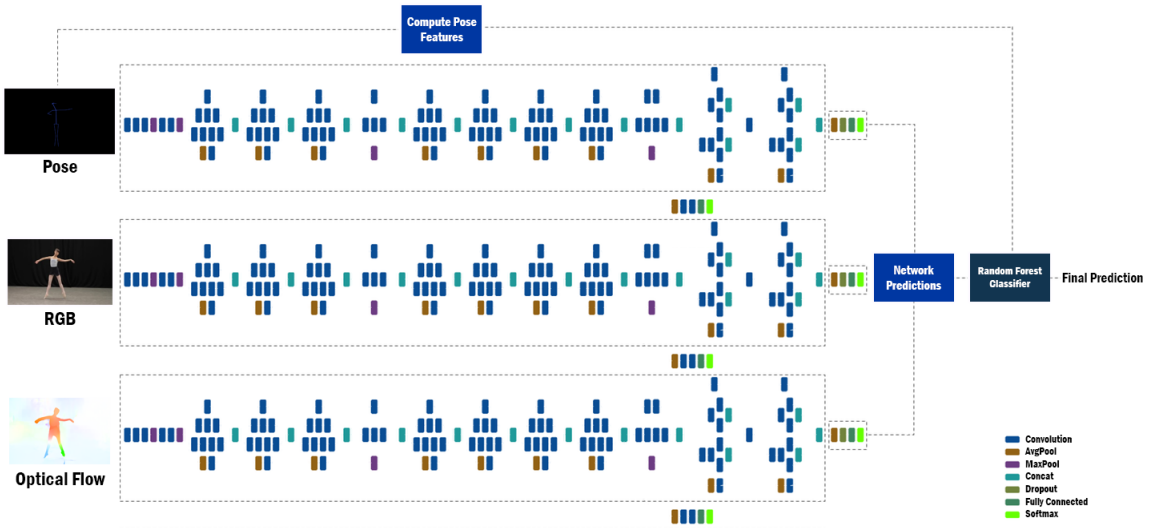

Figure 4.9: Visualization of the Three-Stream + Parameterization Network Architecture. Original visualization of the Inception v3 network was presented here: `https://cloud.google.com/tpu/docs/inception-v3-advanced`

- Elbow (Right and Left)
- Wrist (Right and Left)
- Hip (Right and Left)
- Knee (Right and Left)
- Ankle (Right and Left)

These joints represent a human pose. The changes in position of these joints over time represent human motion. We argue that this motion is vital in improving our understanding of human actions and present a method in which we can leverage this motion alongside state-of-the-art techniques to improve our overall learning.

From here, we can establish some features that encode these joints and more importantly represent the pose of the individual. We encode the following distances in each pose:

- Hip to Elbow (Left and Right)
- Hip to Wrist (Left and Right)

- Shoulder to Wrist (Left and Right)
- Hip to Ankle (Left and Right)
- Wrist Distance (Left to Right)
- Ankle Distance (Left to Right)
- Elbow Distance (Left to Right)
- Polar Distances of Wrist (Left and Right)
- Elbow Angle (Left and Right)

In order to encode each pose into a single feature, we propose an implementation of the Winner-Take-All hashing algorithm to parameterize a human pose [92]. This comparative algorithm encodes the pose by randomly comparing a subset of the distances and using the rank of the maximum value in the subset to generate a hash. The algorithm can actually be used for any feature vector but we believe it is quite well suited for representing a human pose. A single hash therefore represents single human pose in a frame. The number of occurrences of each hash is therefore all of the poses that were seen in the video. This representation can be thought of as a histogram of poses for a specific video. The theory here is that in cases where the class is hard to determine (such as our various types of ballroom dancing), a histogram of the poses seen in a clip improves our understanding of each dance class. This becomes the main set of features we use for learning from a human pose. Given that this algorithm is comparative, we only use the distances to generate each hash and use angles and other parameters as additional features to our histogram of hashes.

We tested a number of parameters for the hashing algorithm and found that $k = 5$ with 3 permutations worked well for our number of parameters. Once we randomly generated a set of permutations, we maintained that set of permutations throughout our experiments (it is important to do this to be able to actually learn from trained features, if your permutations

are different during testing you will experience no learning). The values we used in our work are the following: $[[13, 6, 11, 10, 14], [15, 11, 6, 12, 10], [3, 0, 1, 13, 9]]$.

4.5.5 Movement

In addition to these parameters we encode the number of poses in each frame, we want to incorporate the motion of the pose over time. A simple way of exploring this is to encode a single parameter for motion for each joint, using the following algorithm:

$$m_i = dist(j_i, \frac{\sum_{i=1}^{n} j_i}{n}) \tag{4.2}$$

This computes the distance from the joint i to the centroid of all joints in a given frame. We use the centroid because it is relative to each pose and therefore enables a more stable and accurate representation of change in motion for each joint. After computing this for every pose in every frame, we take the standard deviation of this distance to determine the relative motion of each joint in a dance over time.

4.6 Results & Discussion

In this section we present all of the results we obtained after analyzing our new dataset. All of our results are on a per-video basis (unless stated otherwise) on the same training, testing and validation set. In order to combine frame results, we vote on a per frame basis and pick the majority prediction for a video. Top-k results are presented on a per-frame basis for RGB, Flow and Pose results as shown in Table 4.4.

4.6.1 Top-k Results

As expected, RGB performs best among the three modalities, which is the original input. Flow outperforms pose between 5-10% which we hypothesize is because flow tends to retain more information (sometimes pose fails to recognize any people in the frame). The

Table 4.4: Per-frame Accuracy results for three modalities using the Inception-v3 model pre-trained on ImageNet and fine-tuned on each modality.

Top-K Accuracy	**RGB**	**Flow**	**Pose**
Top-1	47.81%	33.15%	28.01%
Top-3	79.12%	63.99%	54.89%
Top-5	92.86%	78.75%	70.92%

top-5 results demonstrate the model is likely confounding some of the difficult classes, we'll explore this in more depth when analyzing our three-stream models.

4.6.2 Single Stream Network

We compute per-video accuracies for our single-stream networks to compare and assess the raw contribution of each modality. These results mimic the Top-1 results above, with one notable improvement for optical flow which improves by approximately 18%, likely due to very split predictions for certain videos that get correctly classified when assessed on a per-video basis versus a per-frame basis. We see slightly more subdued improvements for pose and RGB, but the rank of each modality remains the same, RGB performs best, followed by flow and then pose (one could speculate this is expected as RGB contains some of the raw information which both pose and flow are directly derived from).

4.6.3 Two-Stream Networks

We combine each of our networks as a two-stream network using late fusion in order to assess their combined performance. Once we combine any two streams on a per-video basis, we see the accuracy converge to approximately 55%, with RGB + Flow achieving the best performance at 56.95%. Interestingly enough, combining flow and pose performs as well as the original modality they were computed from. However, we are able to improve on this by extending this approach to three modalities and finally by aggregating our parameterized approach.

Table 4.5: Per-video Accuracy results for each modality and combined modalities. Accuracy increases steadily from the Top-1 results seen in Table 4.4 because we are assessing on a per-video basis, meaning that non-majority mis-classifications don't count against total accuracy for each video.

Modality	**Accuracy**
Parameterization	39.07%
RGB + Parameterization	54.97%
Flow + Parameterization	51.65%
Pose + Parameterization	38.41%
RGB + Pose + Parameterization	55.63%
RGB + Flow + Parameterization	56.95%
Flow + Pose + Parameterization	54.97%
RGB + Flow + Pose	60.27%
RGB + Flow + Pose + Parameterization	**66.23**%

4.6.4 Three-Stream Network and Parameterization

Lastly, we connect all three streams to achieve the best accuracy using solely deep learning approaches. We see that we are able to categorize approximately 60.27% of the videos correctly. For the three-stream approach, we observe that there are a handful of classes in which we really struggle at classifying. In particular, we struggle with classifying cha, foxtrot, rumba, samba and tango. As we know, these are types of ballroom dancing which have significant visual overlap on a per-frame basis. Adding parameterization to the three-stream network improves accuracy by approximately 6%. However, there are a number of interesting observations to make. Overall, we see improvements in the classification of break dancing, ballet, tango, quickstep among other classes. We see some degradation in performance for cha and foxtrot which is contrasted by improvements in rumba and samba. The biggest outlier is the number of mis-classifications of cha as rumba (7 out of 10 cha videos get mis-classified as rumba). Outside of this, samba gets confused with jive during the three-stream network but some of these mis-classifications are resolved with the added pose parameterization.

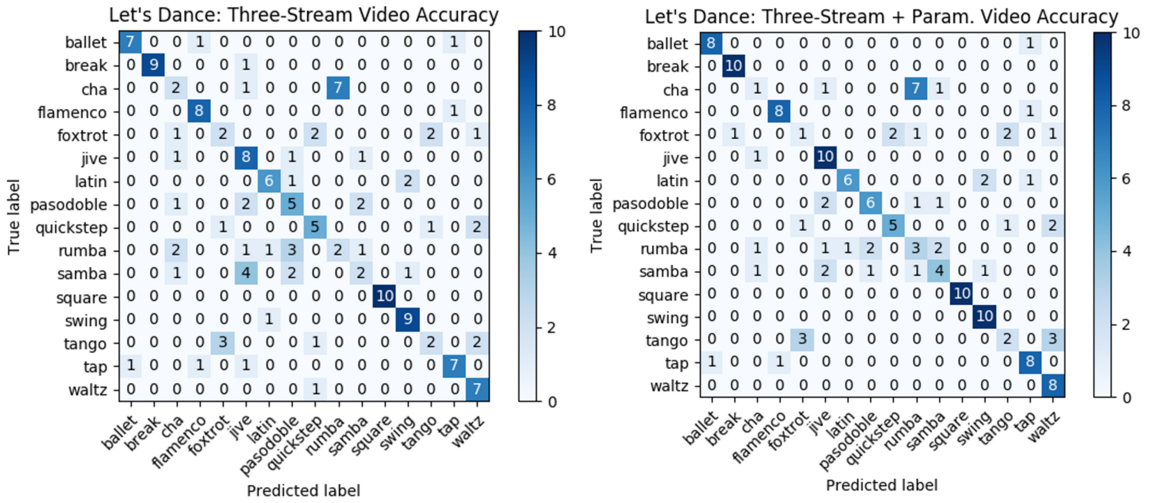

Figure 4.10: **Left**: Three-Stream Class Prediction Confusion Matrix. **Right**: Three-Stream + Parameterization Prediction Confusion Matrix.

4.7 Conclusion

In this chapter we demonstrated the benefit of parameterization in classifying fine-grained human actions. We expanded our existing dataset from 10 to 16 dances, with 5 additional ballroom dances that increase the difficulty of detecting the subtle differences in each dance. From here, we demonstrated that pose representations encode a significant amount of information. Participants in our Mechanical Turk studies were able to identify a variety of activities to high levels of accuracy and were able to understand the action of dancing in a series of examples without any prompt or indication of what the content was about. We learned that computing intuitive parameters alongside training deep neural networks on these representations still has the capability of impacting performance. Although we do believe that a properly designed neuron in a network could learn similar traits to the parameters we hand developed for each of the human poses, we want to note that there is a lot of value in developing metrics that are understandable and representative of the activity. In this work we acknowledge the value of complexity but remain confident that simplicity and intuition carries tremendous value in human understanding.

CHAPTER 5
CONCLUSION

In this thesis we demonstrate the importance of understanding human motion representations and their context for effective action classification. First, we focus on the classification of egocentric images in an everyday context. This classification is achieved by combining the power of deep convolutional neural networks with human intuition in order to fuse a basic parameterization of the data (time of day and color histograms). We also demonstrate the validity of our model by fine-tuning and testing it with two additional participants using one additional day of training data. This highlights that a deep neural network encodes transferable information in the context of daily activity classification. Following that work, we developed the Let's Dance dataset [17]. In this dataset we manually collected 1000 videos of people dancing on YouTube. We hand-selected a 10-second window which actually contained the subjects dancing, and then extracted all of those videos from the web. We then processed all of these videos to generate optical flow representations and human pose information. We then tested common techniques for processing these representations in order to assess the difficulty of classifying these dances. Our best approach required three temporal (3D) neural networks running on each of the representations and achieved an accuracy of 71.60% on the 10 classes. Although this was a promising result, we acknowledge the drawbacks of an incredibly computationally intensive approach to classifying a single action. Therefore, the focus of our work moving forward was in improving the representation and classification of a human pose for achieving comparable accuracy. We demonstrated that a similar, less computationally intensive approach achieved a comparable performance of 70.11%. We also expanded the existing dataset to a total of 1600 videos by adding an additional 6 dancing classes, comprised of 5 Latin ballroom dancing and tap dancing. In addition to this, we improve some of the algorithms behind our optical

flow and pose calculation in order to determine more accurate results and combine these results to achieve state-of-the-art performance for the dataset. Then, we integrate a parameterization of a human pose into our network in order to explicitly encode motion into our learning pipeline. We demonstrate that these explicitly representations introduced through the use of a late-fusion ensemble significantly improve performance and therefore improve the effectiveness of action classification in video. These techniques can be extended to any learning field, especially for problems that lend themselves to an intuitive understanding of the problem space.

Appendices

APPENDIX A

LET'S KEEP DANCING: EXPERIMENT 1: MECHANICAL TURK RESPONSES

Table A.1: Complete List of Results. Many of the repeated results are due to the same user evaluating different videos of the same action. In total, there were 19 participants, on average completing 14 tasks each.

Original Video Class	**Turker Response**
Run	a man running
Gallop sideways	A man is dancing
One-hand wave	A person is saluting
Bend	dance
Jump	na
Jump	left side waallk
Jump in Place	na
Gallop sideways	left side waallk
Gallop sideways	left side waallk
Skip	na
One-hand wave	saluit
One-hand wave	saluit
Two-hand wave	na
Bend	shaking
Jumping Jack	Jumping Exercise
Jump	jump

Table A.1: Continued

Original Video Class	**Turker Response**
Walk	walk
Bend	Just Pick little things from the ground
Jumping Jack	Man doing excercise
Jump	Man moving through jumping with both legs from right side to left side
Jump in Place	Man just jumping in one place
Run	Man Runnning with slowly and slightly
Run	Man Running slightly and slowly
Gallop sideways	Man Jumping with joyfully from right side to left side
Skip	Man Jumping with One leg
Skip	Man jumping with one leg from left side to right side
Walk	Just Walking with long steps
Walk	Man just walking straight from left side to right side
One-hand wave	Man say bye with one hand
One-hand wave	A Man doing salute
Two-hand wave	Man said bye or don't come here
Two-hand wave	One man say bye or don't come here
Run	A figure made up of yellow circles running across screen.
Skip	A figure made up of yellow circles limping.
Bend	A person picking something up with their right hand
Jumping Jack	A person doing jumping jacks
Jump	A person with both arms down hopping forward on both legs towards the right

Table A.1: Continued

Original Video Class	Turker Response
Jump	A person skipping towards the left
Jump	A person hopping forward towards the left with their arms down at their sides
Jump in Place	A person jumping rope
Jump in Place	A person jumping like on a pogo stick
Jump in Place	A person jumping straight up with their legs together and their arms at their side
Run	A person jogging towards the right
Gallop sideways	A person skipping to their right if they are facing me
Gallop sideways	A person facing towards me skipping towards their left with their arms down
Skip	A person hopping towards the left on their right leg
Skip	A person hopping on their right leg and moving forward
Skip	A person hopping on their right leg and moving to the right
Walk	A person wa
Walk	A person walking to the left with their head slightly bowed
One-hand wave	A person with both legs together touching their head with their right arm if they are facing me
One-hand wave	A person touching their head with their right arm if they are facing me
Two-hand wave	A person touching their head from outstretched arms with their legs together
Two-hand wave	A person curling up both arms from an outstretched position and both legs together

Table A.1: Continued

Original Video Class	**Turker Response**
Two-hand wave	A person touching their head with both hands from outstretched arms
Jumping Jack	WORK OUT
Jumping Jack	Animated character doing exercise practice.
Jump	JUMPING
Skip	hopscotch game animation
Walk	Tthe animated character doing walking practice. I think demo demo for robots.
One-hand wave	STAR USING DRESS HUMAN DANCE
Bend	Bending
Bend	Bending
Bend	benting to knee
Jumping Jack	exercise
Jumping Jack	exercise
Jumping Jack	exercise
Jumping Jack	exercise
Jump	jumping on one leg
Jump	jumping on one leg
Jump in Place	jumping
Run	running
Run	running
Gallop sideways	taking long steps
Gallop sideways	taking long steps

Table A.1: Continued

Original Video Class	**Turker Response**
Gallop sideways	taking long steps
Skip	jumping on one leg
Skip	jumping on one leg
Skip	jumping on one leg
Walk	walking
Walk	walking
Walk	walking
Walk	walking
One-hand wave	waving with one arm
One-hand wave	waving with one arm
Two-hand wave	waving with two hands
Two-hand wave	waving with two hands
Two-hand wave	waving with two hands
Two-hand wave	waving with two hands
Bend	a man bending over
Bend	a body bending down
Bend	a body bending down
Bend	a body bending down
Bend	a body bending down
Bend	a body bending down
Jumping Jack	a body doing jumping jacks
Jumping Jack	a man doing jumping jacks with the left arm jerking a bit unnaturally

Table A.1: Continued

Original Video Class	**Turker Response**
Jumping Jack	a man doing jumping jacks
Jump	a body hopping across the screen
Jump	a body jumping across the screen
Jump in Place	a body jumping rope
Run	a body running
Run	a body stumbling as it runs
Run	a body running then slowing down
Run	a body running and then again from the middle of the screen
Run	a body running
Run	a body running
Gallop sideways	a body hopping across the screen sideways
Gallop sideways	a body hopping sideways
Skip	a body skipping
Skip	a body jumping across the screen on one leg
Skip	a body limping
Walk	a body walking
Walk	a body walking
Walk	a body walking
Walk	a body walking
One-hand wave	a body waving
Two-hand wave	a body waving his hands above his head
Two-hand wave	A body moving hands together.

Table A.1: Continued

Original Video Class	Turker Response
Bend	The figure is picking something up from the ground with both arms.
Bend	The figure is picking something up from the ground.
Bend	The figure is picking something up from the ground with the right arm.
Jumping Jack	The figure is doing jumping jacks.
Jumping Jack	The figure is doing jumping jacks.
Jumping Jack	The figure is doing jumping jacks.
Jumping Jack	The figure is doing jumping jacks.
Jump	The figure is hopping on one leg across the screen.
Jump	The figure is hopping across the screen to the left.
Jump	The figure is hopping across the screen.
Jump	The figure is hopping on one leg across the screen.
Jump in Place	The figure is hopping in place.
Jump in Place	The figure is hopping in place.
Jump in Place	The figure is hopping in place.
Run	The figure is running.
Run	The figure is casually running to the left of the screen.
Run	The figure is running to the left of the screen.
Run	The figure is running to the right of the screen.
Run	The figure is casually running.
Gallop sideways	The figure is hopping towards the left.
Gallop sideways	The figure is hopping to the left side.

Table A.1: Continued

Original Video Class	**Turker Response**
Gallop sideways	A figure is side-stepping.
Gallop sideways	The figure is side hopping to the right.
Gallop sideways	A figure is side-stepping to the right.
Gallop sideways	A figure is side-stepping to the right.
Skip	The figure is hopping on the left leg.
Skip	The figure is hopping across with one leg.
Skip	The figure is hopping on one leg.
Skip	The figure is hopping across with one leg.
Skip	The figure is hopping on one leg across the screen.
Walk	The figure is casually walking.
Walk	The figure is casually walking to the left of the screen.
One-hand wave	The figure is waving with the right arm.
One-hand wave	The figure is waving with the right arm.
One-hand wave	The figure is waving.
One-hand wave	The figure is waving with the right arm.
One-hand wave	The figure is waving with the right arm.
Two-hand wave	The figure is waving with both arms.
Two-hand wave	The figure is waving with both arms.
Two-hand wave	The figure is waving with both arms.
Two-hand wave	The figure is waving with both hands.
Bend	A person bending forward.
Bend	A person bending forward.
Jumping Jack	A person doing jumping jacks.

Table A.1: Continued

Original Video Class	Turker Response
Jump in Place	A person jumping rope.
Jump in Place	Jumping
Jump in Place	A person dancing an Irish jig.
Run	A man is running.
Gallop sideways	A person walking sideways.
Walk	A person walking down the street.
Walk	A man is walking
Walk	A person walking.
One-hand wave	A person waving their hand an rubbing their belly.
One-hand wave	A person waving hello.
Two-hand wave	A man is waving his both the hands
Two-hand wave	A man is waving his hand
Jumping Jack	exasais
Jump in Place	playing skipping
Jump in Place	jumping
Jump in Place	jumping
Jump in Place	jumping
Run	running
Skip	one leg jump
Walk	walking
Bend	Reaching down to touch the ground with one arm
Bend	Touching the ground with one hand
Bend	Reaching down to touch the ground with one arm

Table A.1: Continued

Original Video Class	Turker Response
Bend	Reaching down to touch the ground with one arm
Bend	Touching the ground with one hand
Bend	Reaching down to touch the ground with one arm
Bend	Touching the ground with one hand
Bend	Reaching down to touch the ground with one arm
Bend	Reaching down to touch the ground with one arm
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jumping Jack	Jumping jacks
Jump	Jumping around
Jump	Jumping around
Jump	Jumping on both feet
Jump	Jumping
Jump	Hopping on both feet
Jump	Jumping around
Jump	Jumping
Jump	Jumping around

Table A.1: Continued

Original Video Class	**Turker Response**
Jump in Place	Jumping on both feet
Jump in Place	Jumping on both feet
Jump in Place	Jumping on both feet
Jump in Place	Hopping on both feet
Jump in Place	Jumping on both feet
Jump in Place	Hopping
Jump in Place	Jumping up and down
Jump in Place	Hopping
Jump in Place	Hopping
Run	Running
Run	Running slowly
Run	Running slowly
Run	Running
Run	Walking
Run	Running
Run	Running
Run	Running
Run	Running
Run	Running
Gallop sideways	Sliding sideways
Gallop sideways	Jumping sideways
Gallop sideways	Sliding sideways
Gallop sideways	Sliding sideways

Table A.1: Continued

Original Video Class	**Turker Response**
Gallop sideways	Sliding sideways
Gallop sideways	Sliding sideways
Gallop sideways	Sliding sideways
Gallop sideways	Sliding sideways
Gallop sideways	Sliding sideways
Skip	Hopping on one foot
Skip	Jumping on one foot
Skip	Jumping on one foot
Skip	Jumping on one foot
Skip	Hopping on one foot
Skip	Hopping on one foot
Skip	Jumping on one foot
Skip	Hopping on one foot
Skip	Hopping on one foot
Skip	Jumping on one foot
Walk	Walking
Walk	Walking
Walk	Walking
Walk	Walking
Walk	Walking determinedly
Walk	Man walking
Walk	Walking
Walk	Walking

Table A.1: Continued

Original Video Class	**Turker Response**
Walk	Walking fast
Walk	Walking
One-hand wave	Waving one arm
One-hand wave	Waving one arm
One-hand wave	Waving arm
One-hand wave	Waving arm
One-hand wave	Waving arm
One-hand wave	Waving one arm
One-hand wave	Waving arm
One-hand wave	Waving one arm
One-hand wave	Waving one arm
Two-hand wave	Flagging someone down
Two-hand wave	Waving both arms
Two-hand wave	Waving both arms
Two-hand wave	Waving arms
Two-hand wave	Waving both arms
Two-hand wave	Waving arms
Two-hand wave	Waving both arms
Two-hand wave	Waving both arms
Two-hand wave	Waving both arms

APPENDIX B

LET'S KEEP DANCING: EXPERIMENT 2: MECHANICAL TURK RESPONSES

Table B.1: Complete List of Results for Experiment 2. In total, there were 18 participants, on average completing 12 tasks each for a total of 217 descriptions.

Original Video Class	**Turker Response**
jive	This video clip pair dancing and all couples are dancing in fast mode in the floor
jive	One couple as the focus point doing what looks like a quick step. Other couples are behind also dancing in pairs.
waltz	the people are holding and rolling something with their hands
waltz	This video clip single guys are dancing and this is hip hop dance equal dancing in the flooe
pasodoble	This video clip single couple only dancing this floor like a romantic melody theme
tango	All guys are dancing partner dance in the floor like romantic song for this video clip
pasodoble	This video clip all guys are dancing group dance and partner dance also .some peoples watching
foxtrot	This video clip guys dancing like a pair dance and single man dance also included and other peoples are watching the dance

Table B.1: Continued

Original Video Class	Turker Response
rumba	Blue taller man dances with another and then by himself and then lifts up one arm slowly.
rumba	this video clip only one pair dance in melody theme slow motion on the floor
rumba	figures doing the tango
jive	This video clip all guys dancing group dance like a pair dance in the floor
flamenco	This video clip first part couple danced another scenes are single guys dancing the floor
flamenco	four people are dancing
square	This video clip group dancing the floor and guys dancing fast mode
square	Many couples were dancing in a romantic feel.
cha	some people are doing exercise with particular steps
cha	This video clip all guys dancing ballet dance pair changed the dance like a swing
tap	This video clip all guys are dancing in the floor
swing	This video clip only one pair dancing the floor.and surrounding peoples are applauding
swing	Some stick figures are dancing while others watch.
tap	This video clip peoples are dancing line by line like group dancing coordination is perfect
tap	The animated characters are dancing very excitedly.

Table B.1: Continued

Original Video Class	Turker Response
swing	This video clip is partner dance and the couples are dancing romantic music
samba	This video clip pair dancing like a contra dance in the floor
samba	It is a motion video of animated characters. Characters are dancing smartly on the video
break	This video clip single man dancing like a hip hop dance and surrounding peoples are watching the dance
square	First 6 guys are dancing in line and one man added to like a line dancing .surrounding peoples watching the dance
square	several hues of thin stick figures in what appears to be an organized dance.
swing	This video clip all guys are dancing like a group dance and all guys coordination in perfect
tap	best to see
tap	This video clip is dancing like a group dance and the guys are doing line by line and coordination super
tap	It's seven stick figure people dancing, and while they're dancing a few of them change color.
samba	figures disco dancing
samba	This video clip group dancing the floor and single man dancing
samba	Various types of stick figures move around and in and out of frame

Table B.1: Continued

Original Video Class	Turker Response
foxtrot	Solo dancers and couples dancing all over the room. some are dancing fast and some are dancing slow.
foxtrot	figures ballroom dancing
foxtrot	this video clip all guys are dancing the partner dance and single man dance also included
swing	This video clip The single couple only dancing like a swing dance and other guys are applauds
foxtrot	One person in the front dancing around and moving his arms around.
foxtrot	First half single man only dancing then joining the another person like a couple dance in the floor
jive	This video clip only one couple only dancing others not including the floor
cha	This video clip is all guys are dancing partner dance like romantic dance
flamenco	This is dancing stage performance and the guys dancing near the guy using music instrument
tango	This video clip all guys are dancing pair dance and single guy dance also included
quickstep	This video clip single man dancing but all guys are dancing like a solo in the floor
quickstep	A solo dancer dancing in a semi circle followed along by other solo dancers. all the dancers are moving their hips like if dancing to salsa.

Table B.1: Continued

Original Video Class	Turker Response
cha	this video clip is partner dancing and dancing fast mode
tango	This video clip couples are dancing the floor and surrounding peoples are applauding
break	Men with different color LED lights are dancing.
break	this video clip single guys only dancing like a hip hop and other are watching
flamenco	two couple dancing this video clip and this dance is most melody and romantic songs choosed
flamenco	two people in yellow and blue are dancing .
latin	this video clip couples are dancing like a group dance in the floor
pasodoble	This video all guys are dancing couple dancing and single guy dancing also included in the floor
tap	All guys are dancing line by line and group dancing in this video clip
flamenco	This video clip single man performing the stage other guys are using music instruments
flamenco	this video clip this is stage performance and the single guy only dancing other guys are music department
flamenco	great work
samba	This video clip all guys dancing but single man dance own steps in the floor
rumba	This video clip all guys dancing like a couple dance in the floor

Table B.1: Continued

Original Video Class	Turker Response
quickstep	A group of performers wearing LED lighting are dancing around with a few twirling also.
quickstep	All guys are dancing the partner dance and dancing like a fast mode
quickstep	it looks like people dancing together and as singles
tango	In this video all guys are dancing single man dance in the floor .like slow motion dancing
swing	this video clip all guys are dancing like a group dance single man dance in fast mode
tango	Some guys are doing couple dancing and other peoples are single man dancing this video clip
tango	step by step good
tap	All guys are dancing in the floor.this is group dance.beautiful coordination
pasodoble	This video clip ballet dance .and the couples are dancing the flloor and surrounded peoples are sitting watching the dance
pasodoble	Figures sitting in the background while two figures dance together.
swing	This video clip pair dancing romantic themes and surrounding people are enjoy to watching the dance
quickstep	well done work

Table B.1: Continued

Original Video Class	Turker Response
quickstep	Two figures dancing in front then move to back with several other dances and then another dancer moves from left to right in front.
ballet	This video clip single man only dancing like a hip hop dance in the floor
latin	This video clip single couple only dancing this floor like a romantic melody theme
quickstep	Single guy dancing this video clip. one by one dancing the floor ..and show this video
samba	This video clip first half single man dancing then adding one lady pair dancing in fast mode
cha	this video clip all guys are dancing like a partner dance and single man dance also included
cha	multicolored stick figures dancing
rumba	Couples were dancing in the bharatham style.
cha	This video clip only one pair dacing in the floor
cha	figure dancing
square	many colored stick figure walking while the colors on them flicker
square	This video clip couples are dancing and couples exchange dance also going on the floor
flamenco	This video clip group dancing the floor and guys dancing fast mode
flamenco	good effect leg dance

Table B.1: Continued

Original Video Class	Turker Response
waltz	This video clip is guys are dancing like ballet dance .and partner dance also same .gusy looking romantic
waltz	very good couple dance
foxtrot	this video clip all guys are dancing like a partner dance and single man dance also included
break	this video clip one guy only dancing like a hip hop dance and another guy applaud on the floor
break	one in blue giving only hand movements but the yellow one jumping and and giving very fast movements of the body
break	A stick figure is dancing while shapes move around beside them.
quickstep	This video clip all guys dancing couple dance and single man dance also in the fast mode
square	This video clip all guys are dancing partner dance like a swing dancing in the floor
tap	All guys are dancing in line and all are doing hand moments in the floor. in this video clip
samba	This video clip only one pair only dancing in center of the floor and others dancing in themselves in the floor
samba	it looks like some sort of synchronized or choreographed dancing
square	A lone dancer kicking his feet and moving his arms off to the side away from the group. And a group of dancers dancing in a circle and then pairing off.

Table B.1: Continued

Original Video Class	Turker Response
square	this video clip all guys are dancing like a partner dance and single man dance also included
jive	great effect
jive	this is dancing group dance but like a partner dance and guys dance performance is good
samba	This video clip only one couple dancing and surrounding peoples are watching Applauding the floor
ballet	This video clip first half single man only dancing and other lady joining the dance performance solo performance is best part
ballet	figure dancing
ballet	Some figures sitting and while tapping their hands and kicking her feet. Other figures standing and kicking their feet and tapping their hands.
ballet	This video clip single man only dancing and other are watching and applauding
ballet	this video clip 3 members only dancing like that bharatham dance in he floor
tap	this video clip single guy only dancing like a hip hop dance .
break	This video clip single man dancing like a hip hop dance and surrounding peoples are applauding
tango	All guys are dancing like a couple dance with melody themes in this video clip

Table B.1: Continued

Original Video Class	Turker Response
tango	figures dancing
ballet	This video clip like a group dance final stage .and guys are using equipments also
ballet	two group of people are fighting
rumba	this video clip all guys are dancing like a partner dance and single man dance also included
jive	This video clip pair dancing in fast mode and surrounding peoples watching the floor
jive	Two dancers twirl in front around each other and then move slowly to the back with other dancers.
tap	This video clip single man dancing like hip hop in floor and peoples are surrounding applauding and sounded
swing	This video clip only one pair dancing like a ballet dance on the floor
pasodoble	Multiple couples twirling their partners and dancing in in tandem. Other figures sitting outside of the dancing area watching the couples.
pasodoble	This video clip group dancing the floor and guys dancing fast mode
pasodoble	there is a group dance program on the stage
quickstep	This video clip first single man only dancing after lady join with dancing like a ballet dance in this floor
waltz	nice one to see

Table B.1: Continued

Original Video Class	Turker Response
waltz	this video dance like a group dancing but all guys are doing own dance in this floor
jive	this video clip is dancing in the floor and couples are dancing.
jive	Two figures are dancing around while others dance behind them.
pasodoble	The animated characters are dancing so fast and very quickly
pasodoble	This video clip pair dancing romantic themes and surrounding people are enjoy to watching the dance
break	This video clip all guys are dancing the floor only one pair dance
ballet	good to see the dance
ballet	This video clip single man dancing and that guy dancing melody themes
flamenco	Blue person in middle and close standing and moving arms around while a person to the right sitting with others moving in back.
flamenco	This video clip like group dancing and the single man center position of dancing
latin	This video clip single pair dancing the floor like a romantic themes
rumba	Couples are dancing like a contra dance and melody themes are using this performance

Table B.1: Continued

Original Video Class	Turker Response
square	this video clip is couples are dancing like a ballet dance in the floor
square	A group of couples all in a group doing couples dancing. It looks like they are doing different styles.
ballet	Three dancers in a line doing synchronized ballet type moves.
waltz	This video clip pair dancing romantic themes and surrounding people are enjoy to watching the dance
jive	this video clip guys are dancing like a couple dance but single person also danced
quickstep	This video clip is all guys are dancing couple dance and single man dance also included this dance performance
latin	A couple dancing followed by some freestyle solo dancers. Who are interchanging and not all dancing together. Lastly it ends with a couple dancing.
waltz	This video clip all guys are dancing like a partner dance with fast mode in the floor
break	A solo dancer in the middle moving his feet really fast. A few dancers in a line also moving their feet fast. A solo dancer breakdancing their way to the middle.
pasodoble	All guys are dancing like a contra dance couples dancing fast mode in the floor
pasodoble	colourfull dance

Table B.1: Continued

Original Video Class	Turker Response
ballet	This video clip guys are dancing like a group dance and coordination perfect
break	a group of people in different colors dancing with fast movements.
foxtrot	This video clip couples only dancing surrounding peoples are watching and applauding the floor
foxtrot	A group of stick figures are dancing around with one another
latin	This video clip surrounding peoples are dancing single but center of the floor couple are danced
swing	This video clip surrounding peoples are sitting and applauding the center pair dancing
tango	this video clip all guys are dancing like a partner dance and single man dance also included
jive	Different dancers dancing fancy and wildly towards the center of the room.
square	This video clip all guys dancing ballet dance pair changed the dance like aswing
cha	This video clip couples are dancing this floor and peoples are watching the dance
cha	there is a dance competition among couples
waltz	Animated characters showing in the front row are dancing energetically
cha	This video clip pair dancing like a contra dance in the floor

Table B.1: Continued

Original Video Class	Turker Response
foxtrot	This video clip all guys are dancing like a single guy dance others are watching
foxtrot	YES great
rumba	guys are doing swing dancing and the peoples are watching this dance this videoclip
tap	This video clip like a group dance and 4 guys dancing center person hero of the group others coordination is perfect
samba	This video clip all guys are dancing like a partner dance and surrounding peoples are applauding
rumba	This video clip group dance and partner dance and single person also dancing this floor
tango	Randomly the animated characters are dancing so fast.
tango	This video clip all guys dancing like a partner dance and surrounding peoples applauding
rumba	yellow people are dancing well than blue and red.
rumba	Two stick figures are slow dancing together while others dance around them
rumba	This video clip all guys are dancing like a contra dance in the floor
swing	the couple blue and d pink are dancing
swing	A couple dancing with their bodys apart and then twirling and moving in closer then spinning and going a little father apart.

Table B.1: Continued

Original Video Class	Turker Response
swing	This video clip single couple only dancing this floor like a romantic melody theme
break	this video clip single man only dancing like a hip hop dance all guys are applauding
break	some people are participating in a dance completion
quickstep	A group of solo dancers that are dancing separately. One dancer moves into the middle of the dance floor and a few follow him into the center while they dance.
quickstep	few people are dancing and few are watching
quickstep	this video clip all guys are dancing but not a partner dance like a single man dancing the floor
square	This video clip contra dance and the partners are dancing in romantic mood
square	A group of stick figures are walking around grabbing hands.
cha	A couple dancing doing a dance routine with synchronized movements. People sitting behind them watching them dance.
cha	This video clip only one couple dancing and surrounding peoples are watching Applauding the floor
rumba	This video clip all guys are dancing group dance and partner dance
tap	A solo dancer who looks almost looks like he is tap dancing and twirling around. As others sit and stand around him some tapping with him some clapping.

Table B.1: Continued

Original Video Class	Turker Response
tap	This video clip only one man dancing other peoples are applauding watching the floor
foxtrot	This video clip all guys are dancing like a hip hop dance in the floor
samba	this video clip like a group dancing and coordination is perfect
samba	Neon lines in a crude shape of people flicker and appear to look like they are dancing.
foxtrot	figures dancing
foxtrot	this video clip the single pair dancing like a bharatham dance and others single man dancing this foor
samba	This video clip group partner dancing .and guys are doing romantic dancing
samba	Two stick figures are dancing while others move around them
latin	This video clip group dancing the floor and guys dancing fast mode
cha	A group of figures that look like they are doing a fast ballroom routine. Some of the figures are off to the side just watching.
cha	All guys are dancing like a couple dance and this dancing is fast mode
quickstep	This video clip guys dancing like a pair dance and surrounding peoples are applauding

Table B.1: Continued

Original Video Class	Turker Response
quickstep	in a dance program each one demonstrates well
break	figures breakdancing
break	All guys are dancing not a group dance hip hop dancing the floor
quickstep	Couples dancing around some are dancing slow and some are moving very fast. Almost doing a quick step.
quickstep	this video clip all guys are dancing like a partner dance and single man dance also included
samba	This video clip group dance center of the group person single dancing near guys are doing couple dance
latin	This video clip one pair dancing romantic mode in the floor
latin	Laser light stick figures that change color are dancing with each other while colorful, tiny, stick figure people in the distance to the right walk by.
foxtrot	camera pans to right showing dancers in back and then some come from left side closer to camera and then they move to back also.
foxtrot	A few dancers doing a partnered ballroom dance.
foxtrot	This video clip all guys are dancing like a partner dance in the floor
flamenco	This video clip all guys are dancing in the floor but single man danced the floor
flamenco	This video clip single man dancing like a hip hop dance in the floor

Table B.1: Continued

Original Video Class	**Turker Response**
pasodoble	This video clip all guys are dancing own steps not like a group dance in the foor
tango	good
tango	figures doing the tango
tango	group dancing this floor.and surrounding peoples are applauding the group dancing .
swing	This dance only one couple danced like a ballet dance and surrounded people applauding watching this dance
latin	This video clip pair dancing romantic themes in the floor

REFERENCES

[1] E. E. Tripoliti, A. T. Tzallas, M. G. Tsipouras, G. Rigas, P. Bougia, M. Leontiou, S. Konitsiotis, M. Chondrogiorgi, S. Tsouli, and D. I. Fotiadis, "Automatic detection of freezing of gait events in patients with parkinson's disease," *Computer methods and programs in biomedicine*, vol. 110, no. 1, pp. 12–26, 2013.

[2] J. Verghese, R. B. Lipton, C. B. Hall, G. Kuslansky, M. J. Katz, and H. Buschke, "Abnormality of gait as a predictor of non-alzheimer's dementia," *New England Journal of Medicine*, vol. 347, no. 22, pp. 1761–1768, 2002.

[3] M. L. Gleicher and F. Liu, "Re-cinematography: Improving the camerawork of casual video," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 5, no. 1, p. 2, 2008.

[4] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame video stabilization with motion inpainting," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 1150–1163, 2006.

[5] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato, "Sift features tracking for video stabilization," in *14th International Conference on Image Analysis and Processing (ICIAP 2007)*, IEEE, 2007, pp. 825–830.

[6] M. Grundmann, V. Kwatra, and I. Essa, "Auto-directed video stabilization with robust l1 optimal camera paths," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 2011, pp. 225–232.

[7] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 287–295.

[8] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *CVPR 2011*, IEEE, 2011, pp. 3241–3248.

[9] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention prediction in egocentric video using motion and visual saliency," in *Pacific-Rim Symposium on Image and Video Technology*, Springer, 2011, pp. 277–288.

[10] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3216–3223.

[11] X. Ren and C. Gu, "Figure-ground segmentation improves handled object recognition in egocentric video," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010, pp. 3137–3144.

[12] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[13] ——, "Visual motion perception," *Scientific American*, vol. 232, no. 6, pp. 76–89, 1975.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *Computer Vision, 2009 IEEE 12th International Conference on*, IEEE, 2009, pp. 1219–1225.

[16] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[17] D. Castro, S. Hickson, P. Sangkloy, B. Mittal, S. Dai, J. Hays, and I. Essa, "Let's dance: Learning from online dance videos," *arXiv preprint arXiv:1801.07388*, 2018.

[18] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa, "Predicting daily activities from egocentric images using deep learning," in *proceedings of the 2015 ACM International symposium on Wearable Computers*, ACM, 2015, pp. 75–82.

[19] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood, "Sensecam: A retrospective memory aid," in *UbiComp 2006*, Springer, 2006, pp. 177–193.

[20] G. O'Loughlin, S. J. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, and G. D. Warrington, "Using a wearable camera to increase the accuracy of dietary analysis.," *American journal of preventive medicine*, vol. 44, no. 3, pp. 297–301, Mar. 2013.

[21] M Sun, J. D. Fernstrom, W Jia, S. A. Hackworth, N Yao, Y Li, C Li, M. H. Fernstrom, and R. J. Sclabassi, "A wearable electronic system for objective dietary assessment," *Journal of the American Dietetic Association*, vol. 110, no. 1, p. 45, 2010.

[22] G. Marcu, A. K. Dey, and S. Kiesler, "Parent-driven use of wearable cameras for autism support: A field study with families," *Ubicomp 2012*, pp. 401–410, 2012.

[23] P. Kelly, A. Doherty, E. Berry, S. Hodges, A. M. Batterham, and C. Foster, "Can we use digital life-log images to investigate active and sedentary travel behaviour? Results from a pilot study," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 8, no. 1, p. 44, May 2011.

[24] J Kerr, S. J. Marshall, S Godbole, J. Chen, and A Legge, "Using the SenseCam to Improve Classifications of Sedentary Behavior in Free-Living Settings," 2013.

[25] H. Zhang, L. Li, W. Jia, J. D. Fernstrom, R. J. Sclabassi, and M. Sun, "Recognizing physical activity from ego-motion of a camera.," *IEEE EMBS*, pp. 5569–5572, 2010.

[26] W. C. Willett, "Balancing Life-Style and Genomics Research for Disease Prevention," *Science*, vol. 296, no. 5568, pp. 695–698, Apr. 2002.

[27] J Biagioni and J Krumm, "Days of Our Lives: Assessing Day Similarity from Location Traces," *User Modeling*, 2013.

[28] U. Blanke and B. Schiele, "Daily routine recognition through activity spotting," *Location and Context Awareness (LoCA)*, pp. 192–206, 2009.

[29] F.-T. Sun, Y.-T. Yeh, H.-T. Cheng, C. Kuo, and M. L. Griss, "Nonparametric discovery of human routines from sensor data.," *PerCom*, pp. 11–19, 2014.

[30] T. Huỳnh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," *Ubicomp*, 2008.

[31] N Eagle and A Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[32] N. Eagle and A. S. Pentland, "Eigenbehaviors: identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[33] B. P. Clarkson, "Life patterns : structure from wearable sensors," *Thesis (Ph. D.) MIT*, 2005.

[34] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *ICCV*, IEEE, 2011, pp. 407–414.

[35] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, IEEE, 2012, pp. 2847–2854.

[36] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, Springer, 2014, pp. 818–833.

[37] P. Kelly, S. J. Marshall, H. Badland, J. Kerr, M. Oliver, A. R. Doherty, and C. Foster, "An ethical framework for automated, wearable cameras in health behavior research.," *American journal of preventive medicine*, vol. 44, no. 3, pp. 314–319, Mar. 2013.

[38] E. Thomaz, A. Parnami, J. Bidwell, I. A. Essa, and G. D. Abowd, "Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras.," *UbiComp*, pp. 739–748, 2013.

[39] D. H. Nguyen, G Marcu, G. R. Hayes, K. N. Truong, J Scott, M Langheinrich, and C Roduner, "Encountering SenseCam: personal recording technologies in everyday life," pp. 165–174, 2009.

[40] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia, "Privacy behaviors of lifeloggers using wearable cameras," in *ACM International Joint Conference*, 2014, pp. 571–582.

[41] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[42] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014, pp. 675–678.

[45] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, IEEE, 2009, pp. 248–255.

[47] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[48] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *arXiv preprint arXiv:1604.06573*, 2016.

[49] W. Zou, S. Zhu, K. Yu, and A. Y. Ng, "Deep learning of invariant features via simulated fixations in video," in *Advances in Neural Information Processing Systems 25*,

F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 3203–3211.

[50] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, 2013.

[51] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[52] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3218–3226.

[53] G. Gkioxari and J. Malik, "Finding action tubes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 759–768.

[54] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014, pp. 568–576.

[55] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14, ACM, 2014, pp. 97–106, ISBN: 978-1-4503-3063-3.

[56] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[57] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4534–4542.

[58] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," *CoRR*, vol. abs/1502.04681, 2015.

[59] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[60] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*, Springer, 2016, pp. 816–833.

[61] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[62] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *arXiv preprint arXiv:1602.00134*, 2016.

[63] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," 2017.

[64] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[65] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[66] C. Pantofaru, C. Sun, C. Gu, C. Schmid, D. Ross, G. Toderici, J. Malik, R. Sukthankar, S. Vijayanarasimhan, S. Ricco, *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," 2017.

[67] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, IEEE, 2008, pp. 1–8.

[68] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image analysis*, Springer, 2003, pp. 363–370.

[69] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv preprint arXiv:1506.02640*, 2015.

[70] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "Using k-poselets for detecting people and localizing their keypoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3582–3589.

[71] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, 2012, pp. 3178–3185.

[72] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[73] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097–1105.

[74] M. Abadi and A. A. et. al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015.

[75] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[76] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *arXiv preprint arXiv:1507.02159*, 2015.

[77] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *arXiv preprint arXiv:1412.0767*, 2014.

[78] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, Ieee, 2011, pp. 1297–1304.

[79] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[80] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293–1301.

[81] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*, British Machine Vision Association, 2008, pp. 275–1.

[82] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Modeling motion of body parts for action recognition.," in *BMVC*, Citeseer, vol. 11, 2011, pp. 1–12.

[83] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, 2012, pp. 1194–1201.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[85] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[86] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[87] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[88] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[89] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2462–2470.

[90] R. A. Guler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," 2018.

[91] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *null*, IEEE, 2005, pp. 1395–1402.

[92] J. Yagnik, D. Strelow, D. A. Ross, and R.-s. Lin, "The power of comparative reasoning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2011, pp. 2431–2438.

VITA

Daniel Alejandro Castro Chin was born and raised in Panamá City, Panamá. He has two older siblings, Jorge and Beatriz and two loving parents, Jorge and Silvia. He attended the International School of Panama and then went to the Georgia Institute of Technology for his undergraduate and graduate career. In addition to his academic work, Daniel also co-owns a brewery with his brother in Panamá called 2 Oceans Brewing. Daniel is tremendously grateful for the love and support of all of his family, his girlfriend Carley-Beth, and all of his friends which have supported him over the years.